TARTU ÜLIKOOL

MATEMAATIKA-INFORMAATIKATEADUSKOND

MATEMAATILISE STATISTIKA INSTITUUT

Sven Erik Ojavee

Geneetiliste päritolukomponentide määramine mitmemõõtmelise statistika meetodite abil

Bakalaureusetöö (9 EAP)

Juhendajad:

Krista Fischer, PhD

Toomas Haller, PhD

TARTU 2015

Geneetiliste päritolukomponentide määramine mitmemõõtmelise statistika meetodite abil

Käesoleva bakalaureusetöö eesmärk on uurida erinevaid võimalusi indiviidi päritolu kirjeldamiseks geeniandmetel põhinevate tõenäosuslike päritolukomponentide abil. Töö teoreetilises osas antakse ülevaade vajalikest geneetika mõistetest ning kasutatavatest statistilistest meetoditest. Töö praktilises pooles hinnatakse päritolukomponentide skoore peakomponentanalüüsi ning lineaarse diskriminantanalüüsi meetodite kombineerimisel. Tulemuste põhjal leitakse aposterioorsed eri rahvusgruppidesse kuulumise tõenäosused TÜ Eesti Geenivaramu andmebaasist pärinevale valimile. Lisaks uuritakse saadud tulemuste seoseid erinevate fenotüübiliste tunnustega ning samuti teostatakse võrdlused analoogsete tõenäosustega, mis on leitud alternatiivsel meetodil – TÜ Eesti Geenivaramu teadlaste poolt välja töötatud MixFit algoritmi põhjal.

Märksõnad: Päritolu, peakomponentanalüüs, diskriminantanalüüs

Methods of multivariate statistics in the estimation of genetic ancestry components

The aim of this thesis is to examine different options for estimating ancestry-based genetic scores. In the theoretical part an overview of necessary genetic terms and concepts is given followed by the description of the statistical methods that are being used. In the second part of the thesis we calculate genetic scores that are based on a method that combines principal components analysis and linear discriminant analysis. Based on the results, posterior probabilities of belonging to a certain ethnic group are calculated for the data of the Estonian Genome Center. In addition, the associations of the resulting probabilities with different phenotypes, as well as with the results of the new MixFit algorithm are analysed.

Keywords: Ancestry, principal component analysis, discriminant analysis

Sisukord

Sissejuhatus

Inimese rahvusliku päritolu määramine on muutunud üha olulisemaks ülesandeks inimese genoomi uurimisel. See on oluline teave nii personaalmeditsiinis kui ka erinevate ajaloo-alaste või demograafiaga seotud uuringute tarvis. Haiguseriskide hindamisel võimaldavad need tulemused eristada rahvusega seotud elukeskkonnast (sh ka toitumine ja muud eluviisid) tulenevaid riske puhtalt geneetilise taustaga riskikomponentidest. Üks põhilisi viise päritolu määramiseks on üksiknukleotiidsete DNA polümorfismide ehk SNP-de uurimine, sest just nende kohta on hetkel olemas kõige rohkem andmeid.

SNP-id on tekkinud geneetiliste mutatsioonide tulemusena, kus raku paljunemisel on üks nukleotiid asendunud teisega. Kui mutatsioon ei kahjusta olulisel määral organismi elujõulisust, siis antakse see koos DNA-ga edasi järeltulevatele põlvedele. Olukorras, kus erinevad rahvused väga palju ei segune, võivad mutatsioonide esinemissagedused rahvuste kaupa erineda. Kuigi reaalsuses on segunemist siiski toimunud, on siiski tõenäoline, et erinevate rahvuste vahel esineb geneetilisi erinevusi.

Käesoleva töö eesmärk on võrrelda kahe erineva meetodiga arvutatud päritolukomponentide väärtuseid ning uurida saadud komponentide seoseid omakorda mõningate fenotüüpide väärtustega. Ühe meetodina kasutatakse peakomponentanalüüsi ja lineaarset diskriminantanalüüsi rakendamist ning teise meetodina Tartu Ülikooli Eesti geenivaramu vanemteaduri Toomas Halleri ja tema kolleegide poolt välja töötatud MixFit algoritmi.

Töö esimeses peatükis antakse ülevaade peakomponentanalüüsist, lineaarsest diskriminantanalüüsist, MixFit algoritmist ning mõningatest meetoditest, mida kasutatakse hiljem võrdlemiseks. Töö teises peatükis kirjeldatakse peakomponent- ja diskriminantanalüüsi rakendamist ning võrreldakse saadud tulemusi MixFit algoritmi abil saadud tulemuste ja fenotüüpidega. Valdav osa töö käigus tehtud statistilisest analüüsist ja kõik joonised on koostatud statistikapaketiga R, mõne analüüsi jaoks on kasutatud ka statistikapaketti SAS. Kõik töös kasutatud programmid on esitatud lisas 3.

Autor tänab käesoleva bakalaureusetöö juhendajaid Tartu Ülikooli Eesti geenivaramu vanemteadureid Krista Fischerit ja Toomas Hallerit kasulike nõuannete ning huvitava teema püstitamise eest.

1. Ülevaade kirjandusest

1.1 Põhimõisted geneetikast

DNA ehk desoksüribonukleiinhappeks nimetatakse geneetilist informatsiooni kandvat polümeeri, millest koosnevad geenid (Heinaru, 2012, lk 981). DNA koosneb korduvatest alaüksustest ehk nukleotiididest. Iga nukleotiid koosneb kolmest komponendist: fosfaatgrupist, 5-süsinikulisest suhkrust ehk pentoosist ja N-alusest, tsüklilisest lämmastikku sisaldavast ühendist. DNA suhkur on 2-desoksüriboos. DNA sisaldab nelja põhilist lämmastikalust: adeniin (A), guaniin (G), tümiin (T), tsüstosiin (C). (Heinaru, 2012, lk 207)

Kromosoomideks nimetatakse päristuumsetes rakkudes mitoosi või meioosi ajal nähtavaid valkudega kondenseerunud DNA-molekule (Heinaru, 2012, lk 1022). Inimese iga kromosoom on valkude abil kokku pakitud üks lineaarne DNA-molekul (Heinaru, 2012, lk 57-58).

Geeniks nimetatakse pärilikkuse ühikut, mis asub kromosoomi kindlas punktis (lookuses), geen on DNA segment, mis mõjutab mingi tunnuse kujunemist. (Heinaru, 2012, lk 993). Alleeliks nimetatakse kromosoomi lookuses olevat ühte kahest või mitmest alternatiivsest geeniteisendist (Heinaru, 2012, lk 967). Fenotüübiks nimetatakse organismi vaadeldavaid tunnuseid, mis on määratud tema genotüübi ja keskkonnategurite koostoimes (Heinaru, 2012, lk 991).

Üksiknukleotiidseks polümorfismiks (tihti lühendatakse SNP, nimetatakse ka snippideks ehk *single nucleotide polymorphism*) nimetatakse kindlas DNA-punktis oleva üksiku aluspaari vahetusvarieeruvust populatsioonis: üks nukleotiid (A, T, C või G) on asendunud teisega (Heinaru, 2012, lk 1097). Näiteks on ühel indiviidil DNA lõik TACA**G**GATC, ent teisel lõik TACA**A**GATC. Üksiknukleotiidse polümorfismi määramine on inimese geneetilise varieeruvuse üks põhilisi avaldumisviise. SNP-de keskmine sagedus on üks 200-300 aluspaari kohta, millest peaks järelduma, et inimesed on geneetiliselt 99,9% identsed. (Heinaru, 2012, lk 708). Tüüpiliselt on ühel SNP lookusel 2 alleeli (Aaspõllu, 2007).

Faasimiseks (phasing) nimetakse SNP-i alleelide (A, C, T või G) vanemate päritolu määramist. Sisuliselt tähendab faasimine seda, et saadakse teada, milline alleel kuulub millisele kromosoomi koopiale või millised alleelid esinevad koos samas kromosoomis. (ISOGG, 2015)

1.2 Statistilised meetodid

1.2.1 Peakomponentanalüüs

Käesolev alapeatükk põhineb Tartu Ülikooli matemaatilise statistika instituudi dotsent Imbi Traadi mitmemõõtmelise analüüsi loengukonspekti materjalidel (Traat, 2011, lk 2-4). Olgu meil m lähtetunnust X_i, millest tahame konstrueerida uusi tunnuseid. Üldkogumimudeliks on juhuslik vektor $X = (X_1, \dots, X_m)^T$, kusjuures X iseloomustavad keskväärtusvektor EX ja dispersioonimaatriks $\Sigma = E[(X - EX)(X - EX)^T], m \times m.$

Definitsioon. Peakomponendid $P_i, i = 1,2, \dots, m$ on omavahel mittekorreleeritud uued tunnused, mis on esialgsete tunnuste X_i lineaarkombinatsioonid, kusjuures komponendil P_1 on maksimaalne võimalik dispersioon, komponendil P_2 suuruselt järgmine dispersioon jne.

Seega $P_1 = \alpha^T X$, kus $\alpha = (\alpha_1, \dots, \alpha_m)^T$ valitakse nii, et DP_1 on maksimaalne. Ühtlasi tehakse kitsendus kordajate vektori pikkusele, normeeritakse: $\alpha^T \alpha = 1$.

Peakomponentide vektorit tähistame $P = (P_1, \dots, P_m)^T$. Olgu dispersioonimaatriksi tunnusvektori $\vec{X}$ dispersioonimaatriksi Σ omaväärtused

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$$

ning vastavad neile omaväärtustele vastavad omavektorid olgu $\gamma_1, \gamma_2, \dots, \gamma_m$. Tähistame maatriksi, mille veergudeks on omavektorid järgmiselt:

$$\Gamma \coloneqq [\gamma_1 | \gamma_2 | \dots | \gamma_m], m \times m.$$

Diagonaalmaatriksit, mille diagonaalil on omaväärtused, tähistame:

$$\Lambda \coloneqq \mathrm{diag}(\lambda_1, \lambda_2, \dots, \lambda_m), m \times m.$$

On teada, et omavektorid määratakse nii, et nad oleksid ortogonaalsed:

$$\Gamma^T \Gamma = I, \Gamma\Gamma^{\mathrm{T}} = I, \text{ samuti kehtib seos } \Sigma\Gamma = \Gamma\Lambda.$$

Siit järeldub, et $\Gamma^T \Sigma \Gamma = \Lambda, \Sigma = \Gamma \Lambda \Gamma^T$.

On teada, et esimest peakomponenti määravaks kordajate vektoriks $\vec{\alpha}$ on dispersioonimaatriksi Σ suurimale omaväärtusele λ_1 vastav omavektor γ_1.

Teine peakomponent on defineeritud kui $P_2 = \gamma_1^T \vec{X}$. Mittekorreleeritus esimese peakomponendiga on tagatud, sest omavektorid on ortogonaalsed. $DP_2 = \lambda_2$, mis on suurim võimalik, kui kasutame omavektoreid peakomponentide defineerimiseks. Teame maatriksi jälje omadustest, et

$$DX_1 + \cdots + DX_m = \mathrm{tr}(\Sigma) = \lambda_1 + \cdots + \lambda_m.$$

Seega lähtetunnuste kogudispersioon on võrdne peakomponentide kogudispersiooniga, kusjuures iga järgmise peakomponendi dispersioon on maksimaalne võimalik. Kokkuvõttes, valem i-nda peakomponendi defineerimiseks on $P_i = \gamma_i^T X$ ja tema dispersioon on

$$DP_i = \gamma_i^T \Sigma \gamma_i = \lambda_i.$$

Tähtsuse näitaja i-ndale peakomponendile on osakaal

$$\frac{\lambda_i}{\sum_{j=1}^{m} \lambda_j}, \qquad (1.1)$$

mis näitab kui suure osa tunnuste koguvarieeruvusest kirjeldab i-s peakomponent. Väikese osakaaluga peakomponendid jäetakse enamasti analüüsist välja.

Seega kõikide peakomponentide arvutamiseks peame leidma: $P = \Gamma^T X, m \times 1$.

1.2.2 Lineaarne diskriminantanalüüs

Käesolev alapeatükk põhineb raamatul Diskriminantanalüüs (Koskel, Tiit, Arandi, 1998, lk 55-67).

Vaadeldakse k-mõõtmelist tunnusvektorit $X = (X_1, \dots, X_k)^T$. Olgu see vektor mõõdetud g populatsioonis $\Pi^{(h)}, h = 1, \dots, g$, kusjuures osapopulatsioonis h on tunnusvektoril X keskväärtusvektor $\mu^{(h)} = (\mu_1^h, \dots, \mu_k^h)$ ja $\Sigma^{(h)} = \left(\sigma_{ij}^h\right)$ on kovariatsioonimaatriks.

Vaatame juhtu, kus kõik kovariatsioonimaatriksid on võrdsed: $\Sigma^{(h)} = \Sigma, h = 1, \dots, g$. Lisaks eeldame, et Σ on positiivselt määratud, Σ astak on seega k. Iga osapopulatsiooni $\Pi^{(h)}$ valimise tõenäosust juhusliku valiku korral kogupopulatsioonist Π iseloomustab selle osapopulatsiooni tõenäosus, mida tähistatakse π_h.

Ühendatud populatsiooni (Π) keskväärtusvektorit tähistatakse sümboliga $\mu = (\mu_1, \dots, \mu_k)^T$ ning see avaldub $\mu = \sum_{h=1}^{g} \pi_h \mu^{(h)}$. Osapopulatsioonide keskväärtuste $\mu^{(h)}$ kovariatsioonimaatriks avaldub kujul:

$$B = \frac{1}{g} \sum_{h=1}^{g} \pi_h (\mu^{(h)} - \mu)(\mu^{(h)} - \mu)^T.$$

Eeldatakse, et tunnusvektor on klassisiseselt mitmemõõtmelise normaaljaotusega.

Valimit, mis pärineb h-ndast osapopulatsioonist, nimetatakse h-ndaks klassiks mahuga n_h. Seega h-ndasse klassi kuuluvad vaatlused moodustavad andmemaatriksi $X^{(h)}$, milles paiknevad n_h objektil mõõdetud k-mõõtmelise tunnusvektori väärtused. Osapopulatsioonide keskväärtusvektorite $\mu^{(h)}$hinnangud $\bar{x}^{(h)} = (\bar{x}_1^{(h)}, \dots, \bar{x}_k^{(h)})^T$ saab leida standardsel viisil ning neid nimetatakse klassikeskmiseteks. Et eelduse kohaselt kõigis osapopulatsioonides on ühesugune kovariatsioonimaatriks Σ, siis saame selle jaoks ühise hinnangu kõigi klasside kovariatsioonimaatriksite kaalutud keskmisena:

$$S_{(n)} = \frac{1}{n_1 + \cdots + n_g - g} \sum_{h=1}^{g} (n_h - 1) S^{(h)}.$$

Seda valimkovariatsioonimaatriksit nimetatakse klassisiseseks kovariatsioonimaatriksiks. Valimi põhjal saame leida ka hinnangu $\hat{B}$ osapopulatsioonide keskväärtusvektorite hajuvust iseloomustavale klassidevahelisele kovariatsioonimaatriksile B:

$$\hat{B} = \frac{1}{n_1 + \cdots + n_g - g} \sum_{h=1}^{g} (n_h - 1)(\bar{x}^{(h)} - \bar{x})(\bar{x}^{(h)} - \bar{x})^T.$$

Kanoonilise diskriminantanalüüsi idee põhineb R. A. Fisheri ideele projekteerida mitmemõõtmeline tunnusvektor sellistele sihtidele (tunnuse lineaarkombinatsioonidele), mis uuritavaid populatsioone kõige paremini eristavad.

Vaatleme esialgsete tunnuste suvalist lineaarset kombinatsiooni $Y = e^T X$. Osapopulatsioonis $\Pi^{(h)}$ on selle lineaarkombinatsiooni keskväärtus

$$\mu_y^{(h)} = E\left(Y\middle|\Pi^{(h)}\right) = e^T E\left(X\middle|\Pi^{(h)}\right) = e^T \mu^{(h)}$$

ja selle lineaarkombinatsiooni keskväärtus kogu populatsioonis on

$$\mu_y = e^T \mu,$$

kus μ on tunnusvektori X keskväärtusvektor populatsioonis Π. Punktid $\mu^{(h)}$ asuvad kõik ühel sirgel, mille sihi määrab valitud lineaarkombinatsioon $e^T X$. Vastavalt tehtud eeldusele on kõigis osapopulatsioonides ühine kovariatsioonimaatriks, järelikult on ka tunnuse Y dispersioon kõigis osapopulatsioonides võrdne:

$$\sigma_y^2 = e^T \Sigma e.$$

Vaadeldes nüüd osapopulatsioonide teisendatud keskpunkte $\mu_y^{(h)}$ kui võrdtõenäoseid punkte, saame leida nende hajuvust iseloomustava dispersiooni

$$\sigma_B^2 = e^T B e,$$

kus B on osapopulatsioonide keskväärtusvektorite kovariatsioonimaatriks. Kõigi võimalike lineaarkombinatsioonide $Y = e^T X$ hulgast pakub meile huvi leida selline, mis maksimiseeriks osapopulatsioonide keskpunktide dispersiooni σ_B^2, võrrelduna selle lineaarkombinatsiooni Y enese dispersiooniga σ_y^2. Tähistame selle suhte tähega R:

$$R := \frac{e^T B e}{e^T \Sigma e}.$$

Vaja on leida vektori e selline väärtus, mille korral R omandab maksimaalse väärtuse. Normeerime teisendusvektori *e* nii, et oleks rahuldatud $e^T \Sigma e = 1$. Seega tähendab optimaalse vektori e otsimine suuruse $e^T B e$ maksimumi otsimist tingimusel, et $e^T \Sigma e = 1$. On võimalik näidata, et eelkirjeldatud ekstreemumülesanne taandub maatriksi $M = \Sigma^{-1} B$ omaväärtusülesande lahendamisele. Et M ei ole sümmeetriline, taandatakse tavaliselt selle omaväärtusülesande lahendamine sellega seotud maatriksi

$$M^* = \Sigma^{-\frac{1}{2}} B \, \Sigma^{-\frac{1}{2}}$$

omaväärtusülesande lahendamisele. Maatriksi M^*astak olgu s.

Järelikult on maatriksil M^* s positiivset omaväärtust $\lambda_1, \dots, \lambda_s$, mille kohta eeldame, et need on kahanevalt järjestatud, ja sama arv omavektoreid $v^{*1}, \dots, v^{*s}$, kus $v^{*i} = (v_{i1}^*, \dots, v_{is}^*)^T$ ning $i = 1, \dots, s$. Tähistades M^* omaväärtuste maatrikis sümboliga $\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_s)$ ja omavektorite maatriksi sümboliga $V^* = (v_{i1}^*, \dots, v_{is}^*)$, saame kirjutada

$$M^* = V^* \Lambda V^{*T}.$$

Maatriksitel M, M^* on ühised omaväärtused. Maatriksil M on vasakpoolsed ja parempoolsed omavektorid, $M = V \Lambda U^T$, kusjuures

$$M = \Sigma^{-\frac{1}{2}} V^* \Lambda V^{*T} \Sigma^{\frac{1}{2}},$$

millest järelduvad avaldised maatriksi M omavektorite jaoks:

$$\begin{cases} V = \Sigma^{-\frac{1}{2}} V^*, \\ U = \Sigma^{\frac{1}{2}} V^*. \end{cases}$$

Otsitava lineaarkombinatsiooni kordajate vektoriks $e_1 = e$, mis maksimiseerib suhte R, on maatriksi M suurimale omaväärtusele λ_1 vastav vasakpoolne omavektor $v^{(1)}$.

Nimetame lähtetunnuste lineaarfunktsiooni $Y = e_1^T X$ esimeseks diskriminantfunktsiooniks.

Teise, esimesega ristuva diskriminantfunktsiooni määrab maatrikis M teine vasakpoolne omavektor v_2. Nii jätkates on võimalik määrata kokku s tunnusvektori X lineaarkombinatsiooni, mida nimetame diskriminantfunktsioonideks, ja osapopulatsioonide keskväärtusvektoreid täielikult eristada. Seega saime kokku s diskriminantfunktsiooni $d_1, d_2, \dots, d_s$, mille kordajad on määratud maatriksi M vasakpoolsete omavektoritega $v_1, \dots, v_s$. Omavektorid on järjestatud vastavate omaväärtuste kahanemise järgi.

Diskriminantfunktsiooni hindamine

Olgu meil valim tunnusvektori väärtustest. Osapopulatsioonile vastavat valimit nimetatakse klassiks. Tähistame klassi h iseloomustavad suurused:

- Vaatluste arv n_h
- Vaatlustulemused $x_{ij}^{(h)}$, kus $i = 1, \dots, n_h$ näitab objekti ja $j = 1, \dots, k$ tunnuste järjekorranumbrit

Iga objekti iseloomustab k-komponendiline vektor $x_i^{(h)}$.

Iga klassi iseloomustab

- Klassikeskmine (vektor)

$$\bar{x}^{(h)} = \frac{1}{n_h}\sum_{i=1}^{n_h} x_i^{(h)};$$

- Klassi kovariatsioonimaatriks

$$S^{(h)} = \frac{1}{n_h - 1}\sum_{i=1}^{n_h}(x_i^{(h)} - \bar{x}^{(h)})(x_i^{(h)} - \bar{x}^{(h)})^T$$

Kogu andmestikku iseloomustavad

- vaatluste üldarv $n = \sum_{h=1}^{g} n_h$
- üldkeskmine

$$\bar{x} = \frac{1}{n}\sum_{h=1}^{g} n_h \bar{x}^{(h)}$$

- klassikeskmiste hajuvust iseloomustav klassidevaheline kovariatsioonimaatriks

$$\hat{B} = \frac{1}{n_1 + \cdots + n_g - g}\sum_{h=1}^{g}(n_h - 1)(\bar{x}^{(h)} - \bar{x})(\bar{x}^{(h)} - \bar{x})^T.$$

Et oleme eeldanud kovariatsioonimaatriksite võrdsust, saame leida ühise kovariatsioonimaatriksi hinnangu

$$S_{(n)} = \frac{1}{n - g}\sum_{h=1}^{g}(n_h - 1)S^{(h)} = \frac{1}{n - g}W.$$

Maatriks

$$W = \sum_{h=1}^{g}(n_h - 1)S^{(h)}$$

kirjeldab klassisisest hajuvust ehk kõigi klassisiseste hälvete ruutude summasid.

Vaatleme nüüd lähtetunnusvektori X lineaarkombinatsiooni $Y = e^T X$.

Kui teisendusvektor e on teada, saame ka X lineaarfunktsiooni Y parameetritele leida valimi põhjal hinnangud. Tähistame tunnuse Y üldkeskmise ja klasside keskmised vastavalt sümbolitega $\bar{y}, \bar{y}^{(h)}, h = 1, \dots, g$. Tunnuse Y dispersiooni tähiseks olgu sümbol s_y^2.

Diskriminantfunktsioonid

Suhet

$$\hat{R} = \frac{e^T \hat{B} e}{s_y^2}$$

maksimiseeriva empiirilise vektori $\hat{e}$ arvutamiseks tuleb lahendada valimi põhjal arvutatud maatriksi $\widehat{M} = S_{(n)}^{-1}\hat{B}$ omaväärtusülesanne.

Maatriksi $\widehat{M}$ iga vasakpoolne omavektor $\hat{v}^{(j)}$ määrab ühe diskriminantfunktsiooni $e^{(j)}X, j = 1, \dots, s$. Need diskriminantfunktsioonid ei pruugi olla ortogonaalsed, kuid on normeeritud, et kehtib $\hat{e}^T S_{(n)} \hat{e} = I$, kus I on ühikmaatriks. Et eelneva põhjal kehtib võrdus $W = (n-g)S_{(n)}$, kehtib ka seos $S_{(n)}^{-1} = (n-g)W^{-1}$. Järelikult on maatriks $\widetilde{M} \coloneqq W^{-1}\hat{B}$ võrdeline maatriksiga $\widehat{M}$. Vastavalt eelnevatele seostele saame, et $\widetilde{M} = (n-g)^{-1}\widehat{M}$. Sellest aga järeldub, et maatriksite $\widehat{M}$ ja $\widetilde{M}$ omavektorid on vastavalt samasihilised ja seega ka normeeritud omavektorid ühtivad.

Klassikuuluvuse aposterioorse tõenäosuse hindamine

Käesoleva töö puhul pakub huvi aposterioorsete tõenäosuste hindamine. Lähtume eeldusest, et kõigi klasside puhul on tunnusvektor X mitmemõõtmelise normaaljaotusega. Sel juhul saame leida tundmatu objekti Mahalanobise kauguse igast klassikeskmisest $D\left(\bar{x}^{(h)}, x_0\right), h = 1, \dots, g$. Kasutades Mahalanobise kauguse D (Mahalanobise kaugus on defineeritud järgnevalt kahe vektori x, y vahel, kus S on nende kovariatsioonimaatriks:

$$D(x, y) = \sqrt{(x-y)^T S^{-1}(x-y)}$$

) seost F-jaotusega või lähendit normaaljaotuse abil, on võimalik kontrollida nullhüpoteesi selle kohta, kas objekt x_0 kuulub osapopulatsiooni h, kusjuures selle hüpoteesi tõepärasust iseloomustab olulisuse tõenäosus p_h.

Tundmatule objektile kõige lähema klassikeskmise puhul on olulisuse tõenäosus suurim, kaugemate puhul tõenäosus väheneb. Nende andmete põhjal koostatakse diskrimineerimiseeskiri järgnevalt.

Leitakse iga klassi jaoks suhe

$$P_h = \frac{p_h}{\sum_{j=1}^{g} p_j} \quad (1.2)$$

ja nimetatakse suurust P_h klassi h kuulumise aposterioorseks tõenäosuseks.

1.2.3 Spearmani korrelatsioonikordaja

Käesolev alapeatükk põhineb raamatul Statistilise andmetöötluse algõpetus (Parring, Vähi, Käärik, 1997, lk 201-202). Spearmani korrelatsioonikordaja kasutab otseste mõõtmistulemuste asemel nende astakuid, seda kasutatakse tunnuste korral, mis pole normaaljaotusega, ent on parem kui tegemist on pidevate tunnustega.

Kordaja leidmiseks tuleb mõlema tunnuse väärtused järjestada omaette variatsioonritta ja määrata nende astakud. Olgu i-nda objekti tunnuse X väärtuse x_i astakuks s_i, tunnuse Y väärtuse y_i astakuks t_i. Saadud astakuid kasutatakse nagu tavalisi mõõtmistulemusi ja korrelatsioonikordaja leitakse lineaarse korrelatsioonikordaja valemist (Parring, Vähi, Käärik, 1997, lk 187)

$$r = \frac{\sum_{i=1}^{n}(s_i - \bar{s})(t_i - \bar{t})}{\sqrt{\sum_{i=1}^{n}(s_i - \bar{s})^2 \sum_{i=1}^{n}(t_i - \bar{t})^2}}$$

Spearmani korrelatsioonikordaja mõõdab tunnustevahelise monotoonse seose tugevust. Sõltuvust nimetatakse monotoonseks, kui ühe tunnuse keskmine muutus mingis kindlas suunas toob endaga kaasa teise tunnuse muutumise kindlas suunas.

Korrelatsioonikordaja tugevuse hindamiseks kasutatakse järgmiseid piire:

- nõrk seos, kui $|r| \leq 0{,}3$;
- keskmine seos, kui $0{,}3 < |r| < 0{,}7$;
- tugev seos, kui $|r| \geq 0{,}7$. (Parring, Vähi, Käärik, 1997, lk 190)

Kontrollides korraga mitme korrelatsioonikordaja olulisust, on tarvis iga võrdluse puhul kasutada väiksemat olulisuse nivood, et katseviisiline vea tõenäosus ei ületaks mingit väärtust α. Selleks võib kasutada Bonferroni parandust ehk võtta võrdlustes olulisuse nivoodeks $\frac{\alpha}{k}$, kus k on analüüsitava korrelatsioonimaatriksite erinevate elementide arv. (Parring, Vähi, 1995)

1.2.4 Ühefaktorilise dispersioonanalüüsi tasakaalustamata mudel

Käesolev alapeatükk põhineb raamatul Statistilise andmetöötluse algõpetus (Parring, Vähi, Käärik, 1997, lk 270-271). Olgu i-nda valimi maht n_i. Vaatluste koguarv on siis $N = \sum_{i=1}^{k} n_i$, kus k on faktori erinevate tasemete arv. Valimite keskväärtused avalduvad:

$$\bar{y}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$$

ning üldkeskmine:

$$\bar{y}_{..} = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}.$$

F-statistiku arvutamiseks vajalikud hälvete ruutude summad on leitavad järgmistest valemitest:

$$S_A^2 = \sum_{i=1}^{k} n_i \, (\bar{y}_{i.} - \bar{y}_{..})^2,$$

$$S^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2,$$

$$S_y^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2.$$

Juhusliku vea vabadusastmete arvuks on $N - k$.

Tabel 1.2.1 Dispersioonanalüüsi tabel

Varieeruvuse-allikas	Hälvete ruutude summa	Vabadus-astmed	Keskruut	F-suhe
Faktor	S_A^2	$k-1$	$MS_A^2 = \frac{S_A^2}{k-1}$	$F = \frac{MS_A^2}{MS^2}$
Viga	S^2	$N-k$	$MS^2 = \frac{S^2}{N-k}$	
Üldine	S_y^2	$N-1$		

1.2.5 Tukey-Krameri test

Tukey-Krameri test kontrollib võrdlusviisilist viga keskmiste mitmesel võrdlemisel. Test põhineb haarde jaotusel. Algselt kavandatuna eeldas test tasakaalustatud mudelit. Kramer täiendas testi tasakaalustamata mudeli jaoks, esitades mudeli keskmise valimi mahu seosega $\bar{n} = \frac{k}{\frac{1}{n_1}+\cdots+\frac{1}{n_k}}$, kus k on tasemete arv ja n_i on valimimaht i-ndal tasemel ($i = 1, \dots, k$). (Käärik, 2014, lk 61)

Keskväärtuste võrdlemisel tuginetakse Tukey olulise erinevuse statistikule, mis kasutab studentiseeritud haarde kriitilist väärtust $\bar{q}_{\alpha;k,N-k}$, kus α on olulisuse nivoo ja N on vaatluste koguarv.

$$TSD = \bar{q}_{\alpha;k,N-k}\sqrt{\frac{MS^2}{\bar{n}}}$$

Kui $\left|\bar{y}_{i.} - \bar{y}_{j.}\right| \geq TSD$, võetakse vastu sisukas hüpotees,

kui $\left|\bar{y}_{i.} - \bar{y}_{j.}\right| < TSD$, jäädakse nullhüpoteesi juurde. (Parring, Vähi, Käärik, 1997, lk 276-277)

1.3 MixFit algoritm

Käesolev peatükk põhineb Tartu Ülikooli Eesti geenivaramu vanemteadur Toomas Halleri ja tema kolleegide artiklil, mis ei ole veel ilmunud.

Toomas Haller on koos kolleegidega töötanud välja uue analüütilise meetodi, millega on võimalik arvutada indiviidile päritolukomponentide numbrilisi väärtusi. Saadud komponentide väärtuste hinnangud kuuluvad lõiku [0,1] ning esindavad tõenäosust kuivõrd on antud indiviidi esivanemad seotud teatud grupiga (antud olukorras rahvusgrupiga). Komponentide väärtuste hinnangud arvutatakse võrdluspopulatsioonide suhtes. Ühe indiviidi päritolukomponentide summa on 1.

Meetodi muudabki eriliseks tõik, et meetod kombineerib genotüübi andmete faasimise ja arvutused sarnasusmaatriksitega koos mitmemõõtmelise parima sobitamisega. Kirjeldatud lähenemise peamine eelis on meetodi piisav tundlikkus, et suuta eristada juba väikseid geneetilisi erinevusi. Näiteks suudab see eristada muidu üsna sarnaseid eesti ja läti populatsioone.

Arvutusliku teostatavuse hindamiseks kasutati genoomi esindajana kromosoomi 1. Nn „analüütiline toru“ kasutab viimast versiooni SHAPEITi (Delaneau jt, 2014) ja ChromoPainteri (Lawson jt, 2012) tarkvaradest faasimiseks ja järgnevalt arvutab sarnasusmaatriksi. Sellele järgnevalt rakendatakse skripti MixFit (TÜ Eesti Geenivaramu), mis leiab parima sobivuse võrdluspopulatsiooni ning testitud inimese vahel.

„Analüütiline toru“

1. Andmete ettevalmistamine. Võrdlusindiviidide ülegenoomsed andmed koondati ped/map –tüüpi failidesse nii, et iga päritolu võrdlusgrupp oli esindatud sama arvu inimeste poolt. Võrdlusgrupid moodustati inimeste enda teatatud päritolu alusel. Tundmatute inimeste andmed lisati võrdlusfaili lõppu üks korraga.
2. Koondatud genotüübi andmed faasiti programmiga SHAPEIT. Tulemused teisendati formaati IMPUTE2, et jätkata programmiga ChromoPainter.
3. Programmi ChromoPainter kasutati faasitud geeniandmete jagamiseks geneetilisel sarnasusel põhinevatesse lõikudesse. Saadud tükkide loendustulemus on maatriks,

mis loetleb paariviisilise sarnasuse inimeste vahel, võttes aluseks samade genoomitükkide arvu. Iga genoomitükk seatakse alati vastavusse kõige paremini sobivale individuaalsele paarile. See tähendab, et kõik individuaalsed paarid „võistlevad“ genoomitükkide eest. On oluline, et iga tundmatu andmehulk on kombineeritud samade võrdlusandmetega tükkide omistamise protsessis. Iga ChromoPainteri rakendamine andis massiivi (MASSIIV), mis näitab konkreetset individuaalset sarnasust kõikide võrdlusindiviididega ja iseendaga. Sama ChromoPainteri analüüsi korrati ka kõikidele võrdlustele kõikide tundmatute puudumisel nii, et väljastati maatriks (MAATRIKS), mis kirjeldab iga võrdlusindiviidi sarnasust teiste võrdlusindiviididega.

4. Tükkide loendusmaatriksite teisendused. Eelkirjeldatud massiiv sisaldab loendusandmeid ühiste tükkide arvu kohta tundmatute indiviidide ja võrdlusindiviidide vahel. Iga võrdlus kuulub ühte võrdlusgruppi. Kõikide võrdluste ühiste tükkide arv keskmistatakse iga võrdlusgrupi puhul tundmatu jaoks. Tulemusena saadakse indiviidi kirjeldus, mida iseloomustab tema sarnasus iga võrdlusgrupiga tervikuna (leitakse nn „hüpoteetiline keskmine inimene“) ja mitte enam iga võrdlusindiviidiga eraldi. Selline horisontaalne kokkusurumine vähendab veergude arvu maatriksis samale tasemele võrdlusgruppidega. Samasugune horisontaalne kokkusurumine tehakse ka MAATRIKSile. Et MAATRIKS sisaldab samu inimesi nii horisontaalselt kui ka vertikaalselt, surutakse seda samuti sama loogikaga ka vertikaalselt kokku. Saadud maatriksi dimensioonide arv on võrdne võrdlusgruppide arvuga ja iga väärtus esindab keskmist arvu ühistes tükkides kahe võrdlusgrupi vahel. Võrdlusgrupid maatrikis on nüüd esindatud samal viisil kui indiviidid MASSIIVis. MASSIIV ja MAATRIKS normeeritakse üle veergude nii, et iga rea keskmine võrdub ühega. Nende sammudega saadakse geneetilised sarnasusmaatriksid a) tundmatute ja võrdlusgruppide vahel, b) iga võrdlusgrupi ning teiste võrdlusgruppide vahel.
5. MixFit analüüs. MixFit algoritm leiab parima sobivuse MASSIIVI ja MAATRIKSI ridade vahel, et määrata selline võrdluste kombinatsioon, mis kirjeldab kõige paremini tundmatut normaliseeritud keskmise ühise tüki jaotuse kaudu. Võrdluste protsentuaalseid väärtuseid, mis kõige paremini kirjeldavad pärilikkust, nimetataksegi päritolu komponentideks. Võrdluste maksimaalseks arvuks on võetud 3, sest ei saa olla kindel, et enam kui kolme komponendi sobitamine töötab üheselt. Kui tundmatut

kirjeldab kõige paremini vähem kui 3 võrdlust, siis ka vastavat arvu võrdlusi kasutatakse. Kolm paremat päritolu komponenti määratakse, uurides läbi kõik võrdluste kombinatsioonid. (Võrdlusena võib tuua nt värvide lahutamise RGB komponentideks; antud juhul samamoodi lahutatakse „sulam“ protsentuaalseteks komponentideks.) MixFit sobitusprotsess on mitmemõõtmeline sobitusprotsess, kus sarnasust individuaalse ja võrdlusgrupi vahel peetakse maksimaalseks, kui kõikide individuaalsete ja võrdluste vaheliste alamkauguste summa on minimaalne. Alamkaugused on indiviidi päritolu komponentide ja võrdluste vahelised ning neid väljendab grupi-keskmistatud ja normeeritud ühiste genoomitükkide arv. Kahe grupi vaheline kaugus pole defineeritud ainult kui kaugus kindlate geneetiliste päritolu komponentide vahel, vaid kui globaalselt parim sobiv kõikidest päritolu komponentidest. Selline lähenemine võimaldab päritolu komponendid paremini lahutada osadeks, sest kaugused pole ainult lineaarsed mõõdud vaid pigem asukohad mitmemõõtmelises ruumis.

Algoritm

MixFit eraldab kuni kolm võrdluspopulatsiooni, mis sarnanevad kokkuvõttes enim tundmatuga. Alguses on n võrdluspopulatsiooni. Kõiki kombinatsioone (kolm korraga) testitakse teiste suhtes, vähendades järk-järgult nende suhtelisi osakaale kolme võrdlusgrupi segus ning võrreldes tulemusi tundmatuga.

Et võrdlusi muudetakse süstemaatiliselt kolme haaval (ÜHTE muudetakse 0st 1ks, KAHTE muudetakse 1st 0ks ja KOLM on konstantne; seejärel kasutatakse sama loogikat uue väärtuse KOLM korral), muutub sobivus segu ja tundmatu vahel parema ja halvema vahel. Parima sobivuse lokaalsed miinimumid tuvastatakse ning võrdluste osakaalude väärtused salvestatakse. Väärtuseid, mis olid parema 30% miinimumväärtuste seas, hoitakse alles järgnevate sammude jaoks.

Kui kõiki võrdluste kombinatsioone testitakse (juurdekasvuga 0,01), siis kõik võrdluste osakaalude väärtused kõikidest analüüsidest, mis olid 20% paremate seas (tundmatuga sobivuse mõttes), liidetakse referentsi kohta. Igal võrdlusel on väärtus, mis näitab, kui palju oli seda vaja kõikides simulatsioonides, et saavutada parim sobivus. Võrdlused järjestatakse vastavalt nendele skooridele ning kolm kõrgeimat võrdluste skoori ongi tundmatu päritolu

komponendid. Kuna kõik kolm komponenti võisid tulla sõltumatutest simulatsioonidest, tehakse veel üks simulatsioon, et leida sobivaimad osakaalud kolme valitud võrdluse vahel. Selleks viiakse läbi kombinatoorikat kasutav simulatsioon nii, et kõiki kolme võrdluse osakaale testitakse tundmatu suhtes. Sobivamatest 10% väärtustest võetakse lõpptulemuse jaoks aritmeetiline keskmine ning saadaksegi hinnang sellele, millised on sobivaimad osakaalud kolmele võrdlusele.

2. Tartu Ülikooli Geenivaramu kasutuses olevate andmete analüüs

2.1 Andmestike kirjeldus

Kirjeldame edaspidises kolme andmestikku, mida tähistame kui andmestik A, andmestik B ja andmestik C. Andmestik A sisaldab andmeid inimeste kohta, kelle päritolu on juba teada: nad on kas eestlased, lätlased, venelased, lõunasoomlased, põhjasoomlased või rootslased. Andmestik B sisaldab andmeid inimeste kohta Eestist, kuid samas ei ole täpsustatud, milline on nende konkreetne päritolu. Näiteks võib sarnaneda andmestiku B vaatlus hoopis pigem lätlastega, olgugi et geograafiliselt on vaatlus pärit Eestist.

Nii andmestikus A kui ka B on iga vaatluse kohta andmed üle 270 000 SNP-i oleku kohta ehk vaatlused selle kohta, millised alleelid seal esinevad. Andmestikus A on andmeid 568 inimese kohta (100 eestlast, 88 lätlast, 96 venelast, 100 lõunasoomlast, 84 põhjasoomlast, 100 rootslast) ning andmestikus B on andmeid 7 606 inimese kohta.

Andmestik C sisaldab MixFit algoritmiga arvutatud pärilikkusekomponentide väärtuseid ning neid inimesi iseloomustavaid fenotüübilisi tunnuseid andmestiku B indiviidide jaoks. Välja on toodud mitmeid vastavaid fenotüübi väärtuseid, kuid paraku on palju andmeid puudu. Kirjeldatavad fenotüübi väärtused on näiteks sugu, silmavärv, juuksevärv, kaal, pikkus. Ühtlasi on ka infot selle kohta, kui palju inimesed midagi päevas tarbivad, milline on nende haridustase, kui palju neil on lapsi.

Tabel 2.1.1 Näide andmestikust B

	rs2649588	rs2296716	rs2993493	rs2817185	rs4648377
V10544	2	0	1	0	2
V10513	2	0	2	0	1
V11804	1	0	2	0	2
V11476	2	0	2	1	2
V11320	2	0	1	2	2

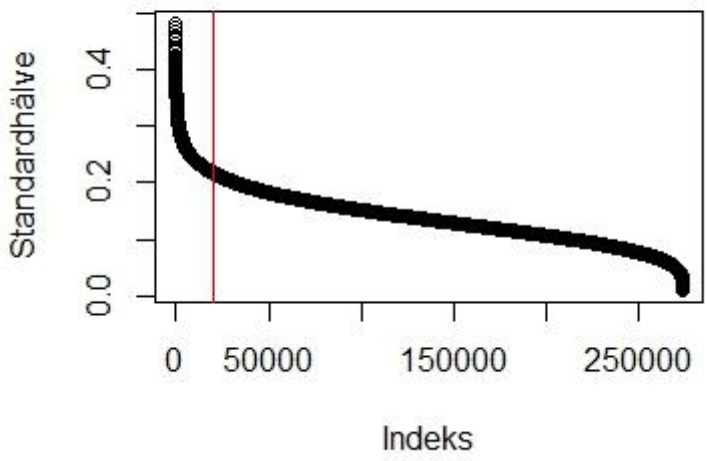
Andmestik A on alamandmestik neist vaatlustest, mida kasutati ka Nelis, Esko jt artikli „Genetic Structure of Europeans: A View from the North-East“ puhul. Selle andmestiku puhul on eestlaste, lätlaste ja venelaste genotüpiseerimine teostatud Eesti Biokeskuses ning andmed soomlaste ja rootslaste kohta on saadud vastavatest riikidest. (Nelis jt, 2009)

2.2 Meetodi kirjeldus

Üldine idee on rakendada SNP-de andmestikule A peakomponentanalüüsi ning analüüsi tulemusena hinnatud peakomponentide maatriksile rakendada diskriminantanalüüsi. Selline metoodika valiti seetõttu, et originaalandmetele pole võimalik otse diskriminantanalüüsi rakendada, sest tunnuste (SNP-de) arv ületab vaatluste (indiviidide) arvu andmestikus. Diskriminantanalüüsi ühe tulemusena on võimalik prognoosida andmestiku B vaatluste aposterioorseid tõenäosuseid ehk tõenäosuseid, et mingi andmestiku B vaatlus kuulub teatavasse andmestiku A poolt kirjeldatud rahvusrühma. Neid tulemusi on võimalik edaspidi juba võrrelda MixFit algoritmi abil saadud tõenäosustega.

Arvutusmahukuse vähendamiseks kasutame andmestikust A vaid 20 000 SNP andmeid, mis on valitud nii, et nende jaotus eri rahvuste vahel oleks võimalikult erinev.

SNP markerite tunnused normeeriti nii, et tunnuse keskmine oleks 0 ja standardhälve 1. Iga markeri jaoks arvutati keskmine iga rahvuse jaoks ning seejärel nende keskmiste standardhälbed. Seejärel valiti 20 000 markerit, mille standardhälbed eri rahvuste vahel olid kõige suuremad. Järgnev peakomponentanalüüs teostati vaid nende 20 000 SNP markeri andmetel. Joonisel 2.2.1 on näha, et kasutatud on ainult punasest vertikaaljoonest vasakul asuvaid väärtusi.



Joonis 2.2.1. SNP markerite rahvuse-spetsiifiliste keskmiste standardhälbed

Edasine analüüs teostati veidi väiksema arvu, 19 585 markeri, andmetega, sest kõigi 20 000 markeri andmeid andmestiku B indiviidide jaoks ei olnud saadaval.

Enne peakomponentanalüüsi rakendamist andmestikule A, normeerime andmestiku A veerud nii, et keskväärtus oleks üks ja standardhälve null. Pärast peakomponentanalüüsi läbiviimist andmestikule A valiti välja 10 esimest peakomponenti, mille alusel leiti (Fisheri) lineaarse diskriminantanalüüsi mudel. Eelnevate tulemuste põhjal andmestikul A arvutati peakomponentide väärtused ka andmestikule B. Selleks normeeriti andmestik B, lahutades igast veerust keskväärtus ja jagades standardhälbega, mida kasutati andmestiku A normeerimiseks enne esialgse peakomponentanalüüsi läbi viimist. Seejärel korrutati saadud andmemaatriks andmestiku A peakomponentanalüüsi poolt väljastatud kordajate α hinnangutega. Saadud peakomponentide väärtuste põhjal leiti aposterioorsed tõenäosused, mis arvutati valemi 1.2 põhjal, et mingi andmestiku B vaatlus kuulub ühte kuue rahvuse klassist.

2.3 Peakomponent- ja diskriminantanalüüsi rakendamine

2.3.1 Peakomponentanalüüsi tulemused

Peakomponentanalüüsi jaoks kasutati 568 inimese 19 585 SNP markeri väärtusi. Enne peakomponentanalüüsi läbiviimist normeeriti SNP markerite väärtused nii, et keskväärtus oli üks ja standardhälve null. Valemi 1.1 abil on võimalik kontrollida, kui suure osa varieeruvusest mingi peakomponent kirjeldab. Osutub, et leitud peakomponendid ei suuda siiski kirjeldada väga suurt osa varieeruvusest, nagu on näha ka järgnevast tabelist juba esimeste peakomponentide põhjal.

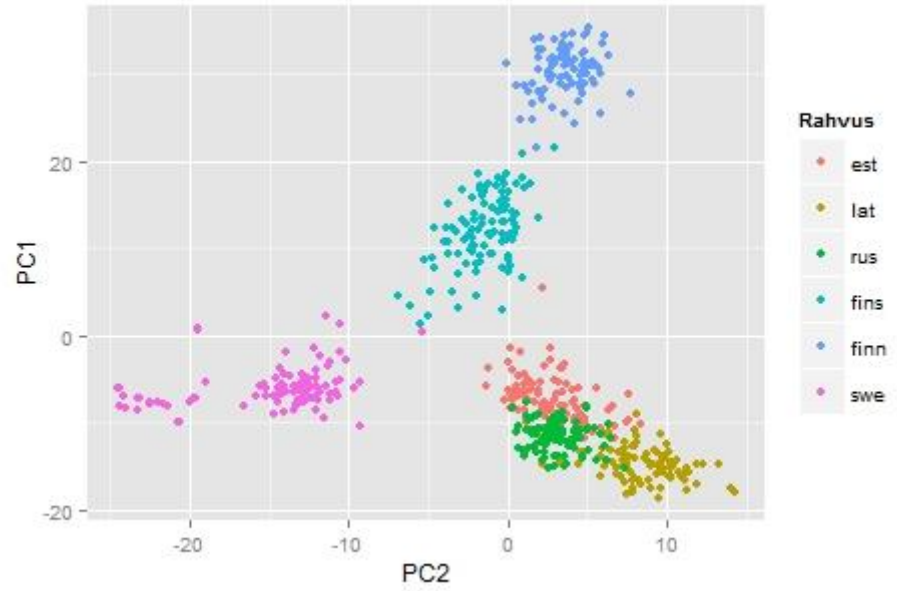
Tabel 2.3.1. Esimese üheksa peakomponendi varieeruvuse kirjeldamine

	PC1	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**	**PC7**	**PC8**	**PC9**
Std h	25,50	13,06	10,23	9,75	8,59	8,57	8,48	8,45	8,38
Osak v	0,0332	0,0087	0,0054	0,0049	0,0038	0,0038	0,0037	0,0037	0,0036
Kum v	0,0332	0,0429	0,0473	0,0521	0,0559	0,0596	0,0633	0,0670	0,0705

Std h – standardhälve, Osak – osakaal koguvarieeruvusest, kum v – kumulatiivne varieeruvus

Võrdlusena võib välja tuua ka eelkirjeldatud artikli, kus uuriti peakomponentanalüüsi abil seoseid eurooplaste geneetilise info ja geograafilise paiknemise vahel. Selles artiklis kirjeldas esimene peakomponent 8,65% ja teine peakomponent 4,68% varieeruvusest. (Nelis jt, 2009) Käesoleva töö tulemus väiksemal andmestikul seega ei suuda kirjeldada nii suurt osa varieeruvusest, mille põhjuseks võib olla see, et kasutati suhteliselt lähedaste rahvuste andmeid ja seega on rahvusgrupi-sisene varieeruvus suhteliselt suur, võrreldes rahvusgruppide-vahelise varieeruvusega. Samas annab joonis 2.3.1, kus ordinaatteljel on

esimene peakomponent ja abstsissteljel teine peakomponent siiski hea geograafilise seose.

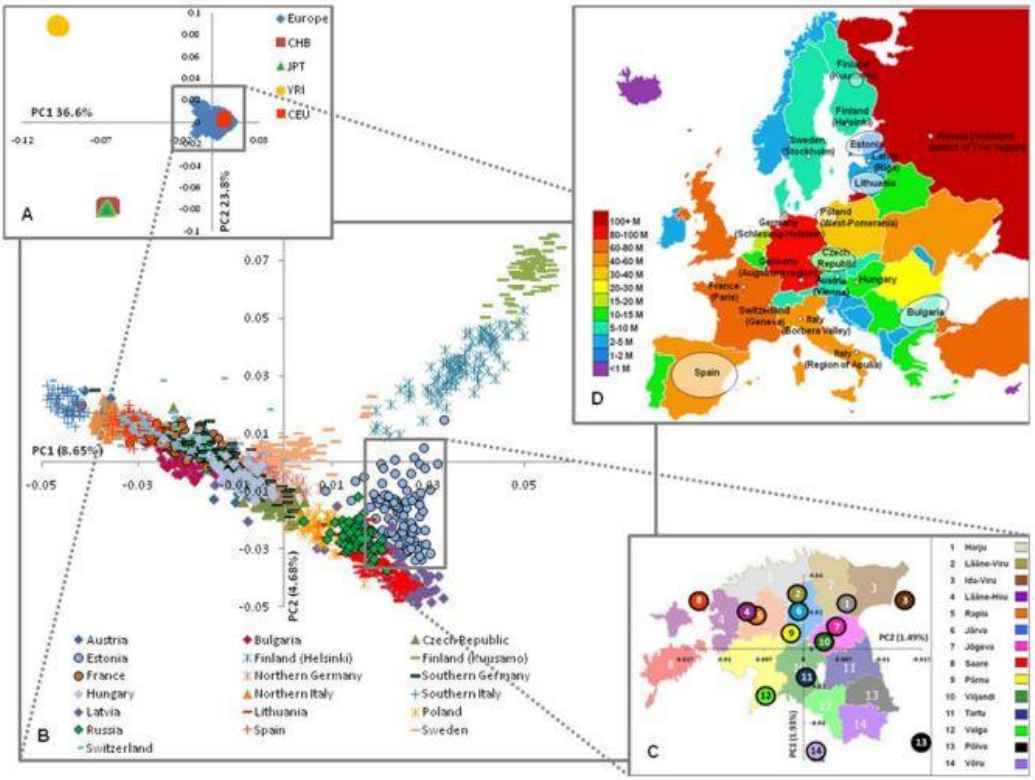


Joonis 2.3.1. Andmestiku A vaatlused kirjeldatud esimese ja teise peakomponendi kaudu

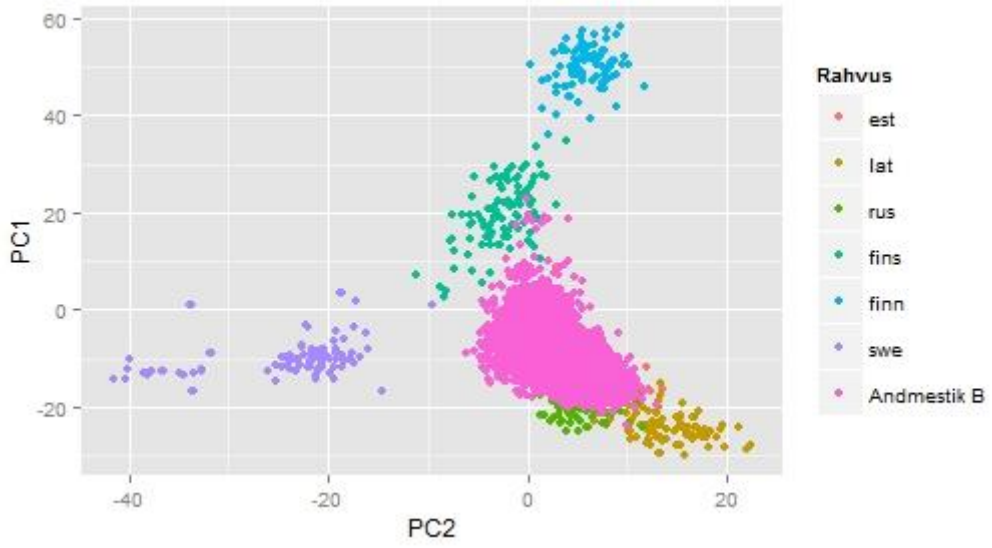
Joonisel 2.3.1 on hästi märgata, et selgesti eristuvad rahvusgrupid on rootslased, lõunasoomlased ja põhjasoomlased. Eestlaste, lätlaste ja venelaste peakomponentide väärtused on märksa sarnasemad, kuid ka nende puhul on võimalik märgata eristuvaid gruppe. Sarnaste jooniste tegemine järgmiste peakomponentide vahel ei andnud interpreteerimiseks väga huvitavaid tulemusi.

Vaatlused näivad olevat samuti seotud rahvuste geograafiliste paiknemisega. Sel juhul esindab esimene peakomponent geneetilise varieeruvuse põhja-lõuna telge ning teine peakomponent geneetilise varieeruvuse ida-lääne telge. Ainus erand selles selgituses on venelaste paiknemine lätlastest pigem lääne pool, kuid ka seda on võimalik seletada vaid Baltikumile pigem lähedal elavate venelaste (Tveri oblast) sattumisega antud valimisse.

Küllalt sarnase seose geograafia ja inimeste genoomi vahel leidsid ka Nelis jt, mida on näha ka jooniselt 2.3.2. Joonise vasakul all paiknevast osast on näha, et sarnaselt eelneva joonisega, on ka sel juhul moodustunud kolmnurk, mille ühes tipus on põhjasoomlased, ühes lätlased ja ühes rootslased. Seega tulemus on üsna sarnane sellega, mida saadi ka eespool. Ühtlasi kinnitab see ka seda, et oli õigustatud suure hulga markerite analüüsi mitte kaasamine, sest artiklis esitatud joonise (joonis 2.3.2) tulemused on saadud, kasutades andmeid enam kui 270 000 SNP-i kohta.



Joonis 2.3.2. Euroopa rahvusgruppide geneetiline paiknemine (Nelis jt, 2009)

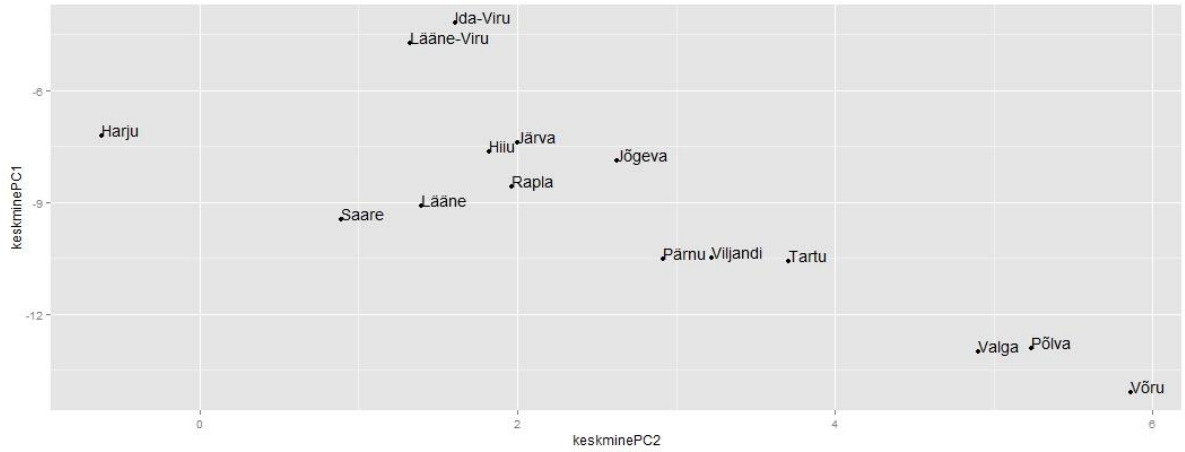


Joonis 2.3.3. Andmestiku A vaatlused, millele on lisatud MixFit algoritmiga uuritud inimeste vastavad tulemused

Joonisel 2.3.3 on kujutatud andmestiku B vaatlused joonisel 2.3.1 esitatud teljestikule.

Jooniselt 2.3.3 on näha, et andmestiku B vaatlused projitseeritakse valdavalt sellesse piirkonda, kus paiknevad enamasti eesti grupi vaatlused, kuid mõned on ka üsna venelaste grupi vaatluste lähedal. Vähem on vaatlusi lätlaste või lõunasoomlaste rühmade juures.

Et enamiku andmestiku B vaatluste juures oli ka välja toodud nende inimeste sünnimaakonnad, on võimalik leida igale maakonnale keskmised esimese ja teise peakomponendi väärtused.



Joonis 2.3.4 Keskmised peakomponentide väärtused andmestikus B maakondade kaupa

Jooniselt on näha, et keskmised peakomponentide väärtused vastavad küllaltki hästi maakondade geograafilisele paiknemisele ning pigem leiab kinnitust see, et esimene peakomponent kirjeldab geneetilise varieeruvuse põhja-lõuna telge ning teine peakomponent ida-lääne telge. Ainus silmapaistev erinevus on seotud Hiiumaaga, ent selle paigutumist üsna Kesk-Eesti maakondade lähedal võib selgitada selle maakonna inimeste väiksem esindatus valimis.

Samas peab rõhutama, et esimesed kaks peakomponenti kirjeldavadki siiski pigem üsna väikese osa (kõigest ligi 4,3%) koguvarieeruvusest ning järgnevaks diskriminantanalüüsiks uuritakse võimalikuks eristamiseks erinevat arvu peakomponente.

2.3.2 Lineaarne diskriminantanalüüs

Lineaarseks diskriminantanalüüsiks on kokku võimalik kasutada kuni 568 peakomponenti. Seega on tarvis otsustada, millise arvu peakomponentide põhjal viiakse läbi edasine analüüs. Edasises on kasutatud võrdlemiseks 2, 5, 10, 50 ja 100 esimest peakomponenti ja neid kõrvutatakse omakorda MixFit algoritmiga saadud tulemustega.

Lineaarse diskriminantanalüüsi puhul eeldatakse, et vaatlused on klassiti mitmemõõtmelisest normaaljaotusest ning kõikide klasside kovariatsioonimaatriksid on võrdsed.

Tabel 2.3.2 Aposterioorsete tõenäosuste hinnangute keskmised sõltuvalt kasutatud peakomponentide arvust

Peak. arv	Osak. varieer.	est	lat	rus	Fins	Finn	swe
2	0,0419	0,8077	0,0043	0,1744	0,0136	0,0000	0,0000
5	0,0559	0,8392	0,0030	0,1531	0,0047	0,0000	0,0000
10	0,0741	0,9417	0,0025	0,0514	0,0043	0,0000	0,0000
50	0,1845	0,9881	0,0027	0,0038	0,0054	0,0000	0,0000
100	0,2990	0,9844	0,0045	0,0045	0,0066	0,0000	0,0000

Peak. arv – peakomponentide arv; Osak. varieer. – antud peakomponentide varieeruvuse osakaal kogu varieeruvusest

Tabelist 2.3.2 on näha, et peakomponentide arvu suurenedes kasvab ka keskmine eesti rahvuse komponendi aposterioorne tõenäosus ning keskmine tõenäosus kuuluda mingisse teise rahvusgruppi kahaneb. Põhjasoomlaste ja rootslaste gruppi kuulumise tõenäosused on hinnatud nulliks iga peakomponentide arvu korral. Selline tulemus on mõnevõrra ka ootuspärane, sest inimesed uuritavas valimis B ongi pärit Eesti aladelt. Samuti on vaadates jooniseid 2.3.1 ja 2.3.3 näha, et kaks peakomponenti määravad suure osa valimist üsna vene ja eesti grupi piirile. Kasutades enam peakomponente, väheneb arvatavasti ka paljude vaatluste võimalus kuuluda vene gruppi.

Tabelis 2.3.3 on näha, kuidas 7606 inimest määrati erinevate rahvusgruppide vahel, kasutades diskriminantanalüüsi teostamiseks eelnevast saadud viit peakomponenti. Inimene määrati sellesse rahvusgruppi, millesse kuulumise aposterioorne tõenäosus oli suurim.

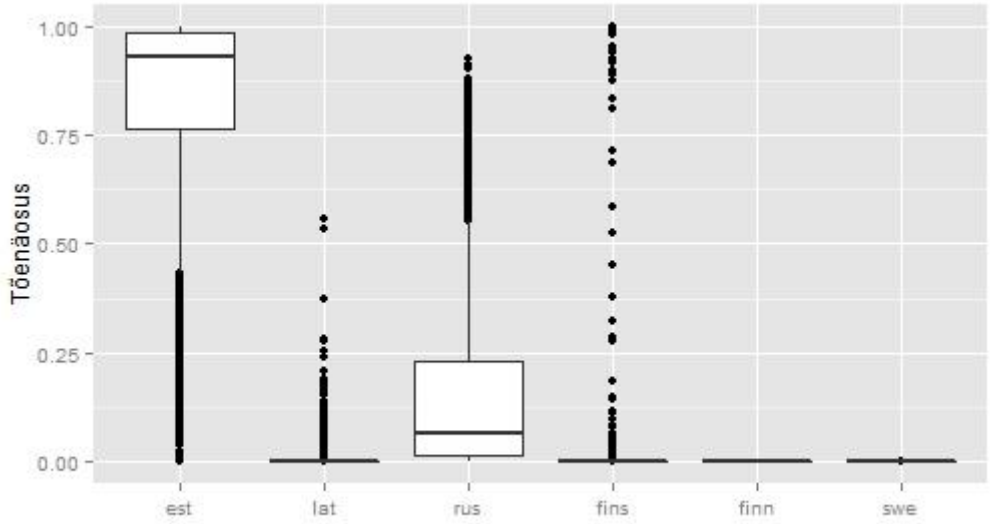
Tabel 2.3.3 Inimeste jaotumine viiele peakomponendile teostatud diskriminantanalüüsi alusel

Rahvusgrupp	Inimesi	Osakaal (%)	Keskmine tõenäosus olla grupis
Eestlased	6921	90,99	0,839
Venelased	648	8,52	0,153
Lõunasoomlased	35	0,46	0,005
Lätlased	2	0,03	0,003
Põhjasoomlased	0	0,00	0,000
Rootslased	0	0,00	0,000

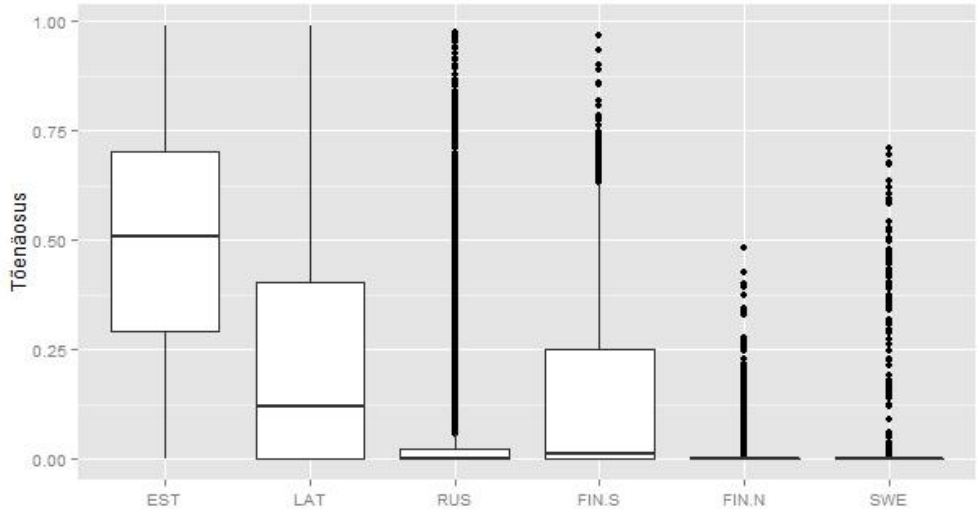
Selgelt on näha, et valdav osa tundmatutest vaatlustest eristatakse eestlastena, mis on ka ootuspärane tulemus, sest ka tõenäosused on suured just eestlaste gruppi kuulumise puhul.

2.4 Tulemuste võrdlemine

Joonistelt 2.4.1 ja 2.4.2 on näha, et MixFit algoritmi abil saadud tulemused erinevad üsna palju peakomponent- ja diskriminantanalüüsi rakendamisel saadud tulemustest. Joonisel 2.4.1 on välja toodud viie peakomponendi põhjal saadud tulemus. Rohkemate peakomponentide kasutamisel on eesti komponendi hinnangud üldiselt veelgi suuremad.



Joonis 2.4.1 Peakomponent- ja diskriminantanalüüsiga leitud tõenäosused kuulumise kohta gruppidesse, kasutatud on viit peakomponenti



Joonis 2.4.2 MixFit algoritmiga leitud tõenäosused kuulumise kohta rahvusgruppidesse

Tabel 2.4.1 MixFit algoritmiga leitud keskmised tõenäosused kuulumise kohta rahvusgruppidesse

	EST	LAT	RUS	FIN.S	FIN.N	SWE
Keskmine	0,5071	0,2223	0,0753	0,1312	0,0052	0,0043

Joonistelt 2.4.1 ja 2.4.2 ning tabelitest 2.3.2 ja 2.4.1 on näha, et peakomponent- ja diskriminantanalüüsi tulemused erinevate peakomponentide arvu korral ning MixFit algoritmiga saadud komponentide väärtused on üsna erinevad. Suur erinevus MixFit algoritmi tulemuste ning peakomponent- ja diskriminantanalüüsi vahel võib tuleneda sellest, et meetodeid on rakendatud mõnevõrra erinevatele lähteandmetele. Kui peakomponent- ja diskriminantanalüüs kasutab SNP-de toorandmeid, siis MixFit algoritmis rakendatakse andmetele eelnevalt ka faasimist ning selle kaudu võib saada lisateavet päritolu kohta, mis mõjutab omakorda lõplikuid hinnanguid päritolukomponentidele.

2.4.1 Korrelatsioonanalüüs erinevate meetoditega saadud päritolukomponentidele

Käesolevas peatükis kirjeldatud korrelatsioonimaatriksid paiknevad lisas 1.

Järgnevalt uuriti, kuidas on omavahel korreleeritud MixFit algoritmi abil saadud tulemused vastavalt 2, 5, 10, 50 ja 100 peakomponenti kasutava diskriminantanalüüsi tulemustega. Selleks kasutati Spearmani korrelatsioonikordajat. Et põhjasoomlaste aposterioorsed tõenäosused tulid kõikide vaatluste ja iga peakomponentide arvu puhul võrdseks nulliga,

jäetakse see grupp edasise analüüsi alt välja. Vastavalt Bonferroni parandusele on iga korrelatsioonikordaja olulisuse kontrollimisel valitud olulisuse nivooks $\frac{0,05}{30} \approx 0,001667$. Järgnevates tabelites (tabelid 2.3.4-2.3.8) on välja toodud rasvases kirjas need korrelatsioonikordajad, mis osutusid oluliseks.

Kõige rohkem pakuvad huvi korrelatsioonikordajad sama rahvusgrupi komponentide hinnangute vahel, mis saadi MixFit algoritmiga või peakomponent- ja diskriminantanalüüsiga. Kõik järgnevalt kirjeldatavad korrelatsioonikordajad vastavate rahvusgruppide komponentide vahel osutusid positiivseteks.

Korrelatsioonikordajad läti komponentide vahel on igas korrelatsioonimaatriksis olulised ning keskmise tugevusega. Kasutatavate peakomponentide arvu suurenedes väheneb ka läti komponentide vaheline korrelatsioon, kui kahe peakomponendiga on see ligi 0,5, siis 100 peakomponendiga on see umbes 0,3. Teisalt on selline vähenemine ka ootuspärane, sest tabel 2.3.2 näitab, et tõenäosused kuuluda eestlaste gruppi kasvavad ning teiste gruppide puhul pigem tõenäosused kahanevad, kui võtta rohkem peakomponente diskriminantanalüüsi sisendiks.

Ka kõik lõunasoome komponentide omavahelised korrelatsioonikordajad osutusid olulisteks. Kahe peakomponendi puhul on lõunasoome komponentide vaheline korrelatsioon keskmise tugevusega, suurema arvu peakomponentide puhul on tegemist nõrkade seostega, ent siiski ei lange korrelatsioonikordaja väärtus alla 0,25.

Nii eesti, rootsi kui ka vene komponentide puhul tulid korrelatsioonikordajad erinevate hinnangute vahel pigem nullilähedased. Kahe peakomponendi puhul osutus korrelatsioonikordaja venekomponentide vahel ebaoluliseks. Peakomponentide arvu kasvades suureneb mõnevõrra ka eestlaste gruppi kuulumise tõenäosuste vaheline korrelatsioonikordaja, kuid jääb ikkagi alla 0,14.

Huvitava tendentsina võib märkida, et peakomponent- ja diskriminantanalüüsiga saadud hinnangud tõenäosuste kohta kuuluda eesti gruppi korreleeruvad märksa paremini MixFit algoritmiga arvutatud tõenäosustega kuuluda lõunasoome gruppi kui tõenäosusega kuuluda eesti gruppi. Need korrelatsioonikordajad vähenevad peakomponentide arvu kasvamisel (5 peakomponendiga on kordaja ligi 0,54, 100 peakomponendiga on kordaja ligi 0,20), ent jäävad siiski iga peakomponentide arvu puhul suuremaks kui korrelatsioonikordajad kahe erineva

metoodikaga arvutatud eestlaste komponentide vahel. Seega näib, et MixFit algoritmiga leitud lõunasoome komponentide väärtustega on pigem seotud peakomponent- ja diskriminantanalüüsiga arvutatud eesti komponentide väärtused.

Analoogiline tulemus eelnevaga kehtib ka lõunasoome ja põhjasoome komponentide vahel. Peakomponent- ja diskriminantanalüüsiga arvutatud lõunasoome komponentide väärtused on tugevamalt korreleeritud MixFit algoritmi abil arvutatud põhjasoomlaste komponendiga kui samal viisil arvutatud lõunasoome komponentidega. Sel juhul on kirjeldatud lõunasoome ja põhjasoome komponentide vaheline korrelatsioonikordaja 0,5 lähedal, seega on tegemist keskmise tugevusega seosega.

Korrelatsioonikordajate uurimise põhjal ei ole võimalik käesoleval juhul järeldada märkimisväärselt tugevaid seoseid MixFit algoritmi ning peakomponent- ja diskriminantanalüüsi tulemuste vahel. Korrelatsioonikordajad erinevate meetoditega leitud sama rahvuse komponentide vahel olid pigem väiksed, vaid läti komponendi puhul oli võimalik täheldada keskmise tugevusega seost.

Samuti on võimalik, et eestlaste, lõunasoomlaste ja põhjasoomlaste tulemused on nii-öelda nihkes. Peakomponent- ja diskriminantanalüüsiga saadud eesti ja lõunasoome komponendid on tugevamalt korreleeritud vastavalt MixFit algoritmiga saadud lõunasoome ja põhjasoome komponentidega.

2.4.2 Seosed fenotüüpide ja päritolukomponentide vahel

Käesolevas peatükis kirjeldatavad tabelid paiknevad lisas 2.

Järgnevalt uuriti kolme fenotüübitunnust: pikkus, silmade värv ja loomulik juuksevärv. Fenotüüpide seoseid vaadati neljal erineval juhul saadud päritolukomponentidega: MixFit algoritmiga saadud komponendid, 2 peakomponendile tehtud diskriminantanalüüsiga saadud komponendid, 5 peakomponendile tehtud diskriminantanalüüsiga saadud komponendid ning 50 peakomponendile tehtud diskriminantanalüüsiga saadud komponendid. Sellised peakomponentide arvud valiti, et vaadelda võimalikult erinevaid juhtumeid, ent samas mitte korrata sarnaseid tulemusi. Et pärast peakomponent- ja diskriminantanalüüside teostamist osutusid põhjasoome komponendi väärtused nullideks iga inimese puhul ning ka rootslaste

komponendi puhul leidusid vaid üksikud nullist erinevad väärtused, on jäetud põhjasoome ja rootsi komponendid järgnenud analüüsidest välja.

Pikkuse puhul vaadati Pearsoni korrelatsioonikordajaid päritolukomponendi väärtuste ning indiviidi pikkust. Silmavärvi ning juuksevärvi puhul teostati (tasakaalustamata) ühefaktorilised dispersioonanalüüsid, kus faktor oli vastavalt kas silma- või juuksevärv, faktori tasemeteks erinevad silma- või juuksevärvi toonid ning uuritavaks tunnuseks mingi päritolukomponent. Oluliste mudelite tekkimisel kontrolliti Tukey-Krameri testiga, milliste värvide keskmised erinevad omakorda oluliselt.

Pikkus

Et iga vaatluse tabeli puhul vaatleme korraga nelja korrelatsioonikordajat, on olulisuse nivooks valitud $\frac{0,05}{4} = 0,0125$. Kahe peakomponendi korral on kõik kordajad olulised, viie korral on olulised kordajad eesti, läti ja vene komponentide puhul, 50 korral on oluline ainult läti komponendi kordaja kordaja. MixFit algoritmi korral on oluline ainult eestlaste kordaja.

Tabelitest on näha, et oluliste korrelatsioonikordajate puhul on kordajad eesti komponendi ja pikkuse vahel positiivsed ning kordajad teiste rahvuste ja pikkuse vahel negatiivsed. Seega suurem pikkus näib olevat seotud suurema eesti komponendi väärtusega olenemata komponendi arvutamise meetodist. Paraku on korrelatsioonikordajate väärtused väga lähedased nullile, mistõttu võib väita, et seos pikkuste ja käesolevate päritolukomponentide vahel on väga nõrk.

Koostame järgnevalt tabelid (tabelid 2.4.1-2.4.5), et kokku võtta analüüsi tulemused. Kui tabelisse on kirjutatud mingi rahvuse tunnus (est, lat, …), siis järelikult on selle rahvuse komponentide väärtused juuksevärvuseti erinevad. Rida näitab värvust, mille korral on vastava komponendi väärtus väiksem, ning veerg näitab värvust, mille korral see on suurem. Näiteks on tabelis 2.4.1 vene komponendi keskmine väärtus blondidel oluliselt madalam vene komponendi keskmisest väärtusest mustade juustega inimeste omast.

Juuksevärv

Dispersioonanalüüsis kasutati faktori tasemetena nelja juuksevärvi: blond, must, pruun, punane.

Tabel 2.4.2 Olulised keskmiste erinevused värvuste vahel kahele peakomponendile teostatud diskriminantanalüüsi korral

LDA2	Blond	Must	Pruun	Punane
Blond		rus	rus	
Must				
Pruun	fins			
Punane				

Tabel 2.4.3 Olulised keskmiste erinevused värvuste esinemissageduste vahel viiele peakomponendile teostatud diskriminantanalüüsi korral

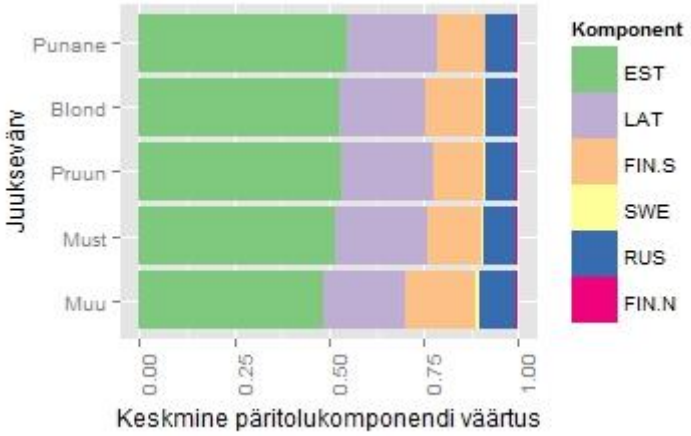
LDA5	Blond	Must	Pruun	Punane
Blond		rus	rus	
Must	est			
Pruun	est			
Punane		rus		

50 peakomponendile teostatud diskriminantanalüüsi korral ei leidunud ühtki olulist erinevust.

Tabel 2.4.4 Olulised keskmiste erinevused värvuste vahel MixFit algoritmi tulemuste korral

MixFit	Blond	Must	Pruun	Punane
Blond				
Must				
Pruun	FIN.S			
Punane				

Joonis 2.4.3 kinnitab tabeli 2.4.4 tulemust. On näha, et keskmine lõunasoome komponendi väärtus on mõnevõrra suurem just blondide hulgas.



Joonis 2.4.3 MixFit päritolukomponentide väärtused juuksevärvide kaupa

Nii MixFit algoritmiga kui ka kahele peakomponendile tehtud diskriminantanalüüsiga saadud lõunasoome komponendi keskmine väärtus on blondidel suurem kui pruunide juustega isikutel. Nii viiele kui ka kahele peakomponendile tehtud diskriminantanalüüsiga saadud vene komponendi keskmine väärtus on blondidel väiksem kui pruunide või mustade juustega isikutel. Peatükist 2.4.1 selgus, et viiele peakomponendile tehtud diskriminantanalüüsiga arvutatud eesti komponent on keskmise tugevusega korreleeritud MixFit algoritmiga arvutatud lõunasoome komponendiga. Seega võib oletada, et MixFit algoritmi ja kahele peakomponendile teostatud diskriminantanalüüsi keskmise lõunasoome komponendi suurem väärtus blondidel kui pruunide juustega inimestel on sarnane tulemus viiele peakomponendile teostatud diskriminantanalüüsi keskmise eesti komponendi suurema väärtusega blondidel kui pruunide juustega inimestel.

Silmavärv

Dispersioonanalüüsis kasutati faktori tasemetena nelja silmavärvi: hall, pruun, roheline, sinine.

Olulisi erinevusi leidus ainult kahele ja viiele peakomponendile teostatud diskriminantanalüüsi tulemustes. 50 peakomponendi või MixFit algoritmi kasutamine ei toonud välja, et

silmavärvuseti leiduks erinevusi keskmiste päritolukomponentide vahel. Olulised erinevused on välja toodud tabelites 2.4.5 ning 2.4.6.

Tabel 2.4.5 Olulised keskmiste erinevused silmavärvuste vahel kahele peakomponendile teostatud diskriminantanalüüsi korral

LDA2	Hall	Pruun	Roheline	Sinine
Hall		Est	Est	est
Pruun	Lat			
Roheline	Rus			
Sinine	Rus			

Tabel 2.4.6 Olulised keskmiste erinevused silmavärvuste vahel viiele peakomponendile teostatud diskriminantanalüüsi korral

LDA5	Hall	Pruun	Roheline	Sinine
Hall				
Pruun	lat			
Roheline				
Sinine				

Silmavärvide puhul olid tulemused erinevate meetoditega arvutatud üsna erinevad, sest statistiliselt olulisi erinevusi erinevate värvide vahel on vähe. Kokkulangev tulemus kaht või viit peakomponenti kasutava diskriminantanalüüsi puhul oli see, et keskmine läti komponent oli hallide silmadega inimestel kõrgem kui pruunide silmadega inimestel. Kahe peakomponendiga juhul oli eesti komponent oluliselt väiksem hallide silmadega inimestel võrreldes kõikide teiste silmavärvidega.

Erinevate juuksevärvide korral leidus enim olulisi erinevusi keskmiste vahel, kokku 5, viiele peakomponendile tehtud diskriminantanalüüsi tulemuste puhul. Kahele peakomponendile tehtud diskriminantanalüüsi puhul leidus selliseid olulisi erinevusi 3 ning MixFit algoritmi tulemuste korral leidus selliseid erinevusi 1. Erinevate silmavärvide korral leidus enim olulisi erinevusi kahele peakomponendile tehtud diskriminantanalüüsi tulemuste puhul, kus olulisi erinevusi leiti 6. Viiele peakomponendile tehtud diskriminantanalüüsi tulemuste põhjal oli võimalik leida 1 paar, mille keskmised erinevad värvuseti oluliselt.

Kokkuvõte

Käesoleva bakalaureusetöö eesmärk oli uurida kahte erinevat meetodit päritolukomponentide hindamiseks geneetilise informatsiooni alusel ja seejärel võrrelda meetoditega saadud komponentide väärtuseid teatavate fenotüübiväärtustega ning teineteise vahel. Töös kirjeldati päritolukomponentide hindamiseks kahte meetodit: TÜ Eesti Geenivaramu teadlaste poolt välja töötatud algoritmi MixFit ning peakomponentanalüüsi ja lineaarse diskriminantanalüüsi järjest rakendamine.

Osutus, et erinevate meetoditega arvutatud päritolukomponentide väärtused erinevad omavahel üsna palju. Näiteks on keskmine eesti komponendi väärtus MixFit algoritmi tulemuste korral TÜ Eesti Geenivaramu valimis ligi 0,51, kahe peakomponendiga tehtud diskriminantanalüüsi tulemuste korral ligi 0,81 ning saja peakomponendiga tehtud diskriminantanalüüsi tulemuste korral ligi 0,98. Suured erinevused on ka teiste rahvusgruppide vahel.

Ehkki päritolukomponentide väärtused erinevate meetodite puhul ei ole väga sarnased, on korrelatsioonikordajad sama rahvusgrupi komponentide MixFit hinnangute ning peakomponent- ja diskriminantanalüüsi vahel positiivsed ja statistiliselt olulised. Tähelepanuväärne on ka asjaolu, et diskriminantanalüüsidega leitud eesti komponendid korreleeruvad tugevamini MixFit algoritmiga leitud lõunasoome komponentidega ning diskriminantanalüüsidega leitud lõunasoome komponendid korreleeruvad tugevamini MixFit algoritmiga leitud põhjasoome komponentidega.

Uuriti ka erinevate meetoditega leitud päritolukomponentide seoseid fenotüüpidega nagu pikkus, juuksevärv ja silmavärv. Korrelatsioonikordajad pikkuste ning erinevate meetoditega arvutatud päritolukomponentide vahel tulid enamasti küll olulised, kuid samas olid need nullile väga lähedal. Seega ei saa väita, et mingisuguse meetodiga saadud tulemused oleksid olnud tugevalt korreleeritud pikkusega. Kõige rohkem olulisi erinevusi juuksevärvide kaupa leitud päritolukomponentide keskmiste vahel tekkis viiele peakomponendile teostatud diskriminantanalüüsi puhul ja kõige rohkem olulisi erinevusi silmavärvide kaupa leitud

päritolukomponentide keskmiste vahel tekkis kahele peakomponendile teostatud diskriminantanalüüsi puhul.

Seega juukse- või silmavärviga on paremini seotud komponendid, mis on arvutatud diskriminantanalüüsi abil, kasutades selleks kahte või viit peakomponenti. Kuigi mõningate fenotüübi väärtustega on paremini seotud vähesele arvule peakomponentidele tehtud diskriminantanalüüsi tulemused, ei ole sellest veel võimalik järeldada, kumb meetod töötab paremini. Siiski tundub käesoleva töö põhjal, et peakomponent- ja diskriminantanalüüsi kombineerimine annab häid tulemusi ja edaspidi võiks uurida, kuidas töötab see meetod, kui teda rakendada juba faasitud andmetele. Erinevate meetodite otsesemaks võrdluseks oleks edaspidi hea kasutada ka andmeid isikute kohta, kelle päritolu mitme põlvkonna lõikes on hästi teada.

Viited

Aaspõllu, A., 2007. Pärilikkusaine DNA eristab meid teistest ja üksteisest, *Horisont*, [online] Kättesaadav: http://www.horisont.ee/node/205 [Vaadatud 08. 04. 2015].

Delaneau, O., Zagury, J.-F.,Marchini, J., 2014. *SHAPEIT veebilehekülg.* [online] Kättesaadav: https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html [Vaadatud 29.04.2015].

Haller, T. jt, ilmumisel. Methodology for computing ancestry-related genetic scores at the individual level and its application to the Estonian and Finnish population studies.

Heinaru, A., 2012. *Geneetika.* Tartu Ülikooli Kirjastus.

International Society of Genetic Genealogy Wiki, *Phasing*. [online] Kättesaadav: http://www.isogg.org/wiki/Phasing [Vaadatud 08.04.2015].

Koskel, S., Tiit, E.-M., Arandi, P., 1998. *Diskriminantanalüüs.* Tartu Ülikooli Kirjastus.

Käärik, E., 2014. *Andmeanalüüs II. Loengukonspekt.*

Lawson,D., Hellenthal, G., Falush, D., Myers, S., 2012. *Chromopainter veebilehekülg.* [online] Kättesaadav: http://www.paintmychromosomes.com/ [Vaadatud 29.04.2015].

Nelis, M, Esko, T. jt, 2009. Genetic Structure of Europeans: A View from the North-East. *Plos One,*[online]. Kättesaadav: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0005472 [Vaadatud 09.04.2015].

Parring, A.-M., Vähi, M., Käärik, E., 1997. *Statistilise andmetöötluse algõpetus.* Tartu Ülikooli Kirjastus.

Parring, A.-M., Vähi, M., 1995. Korrelatsioonimaatriksi ohtlikud olulisuse tõenäosused, *Eesti statistikaseltsi teabevihik nr 5,* [online] Kättesaadav: http://www-1.ms.ut.ee/ess/Failid/Teabevihud/ESSteabevihik5.pdf [Vaadatud 26.04.2015]

Traat, I., 2011. *Mitmemõõtmeline analüüs. Loengukonspekt.*

Lisad

Lisa 1. Korrelatsioonimaatriksid erinevate meetoditega saadud päritolukomponentide kohta

Tabelites L1.1-L1.5 on rasvases kirjas välja toodud korrelatsioonikordajad, mis osutusid oluliseks. Olulisuse nivooks on valitud $0{,}05$.

Tabel L1.1 Spearmani korrelatsioonikordajad MixFit algoritmi ja peakomponentanalüüs+diskriminantanalüüsi tulemuste vahel, 2 peakomponenti

	EST	LAT	RUS	FIN.S	SWE	FIN.N
est	**0,0874**	**-0,4044**	0,0022	**0,4986**	**0,0828**	**0,1389**
lat	**-0,0752**	**0,5014**	0,0017	**-0,5528**	**-0,1070**	**-0,2832**
rus	**-0,0653**	**0,4700**	0,0329	**-0,5833**	**-0,0794**	**-0,3435**
fins	**-0,0538**	**-0,2812**	**-0,0870**	**0,4064**	0,0226	**0,4930**
swe	**-0,0421**	**-0,1204**	0,0043	**0,0897**	**0,1066**	**0,0629**

Tabel L1.2 Spearmani korrelatsioonikordajad MixFit algoritmi ja peakomponentanalüüs+diskriminantanalüüsi tulemuste vahel, 5 peakomponenti

	EST	LAT	RUS	FIN.S	SWE	FIN.N
est	**0,0782**	**-0,4223**	-0,0313	**0,5405**	**0,0828**	**0,2513**
lat	**-0,0889**	**0,4858**	0,0045	**-0,5415**	**-0,1030**	**-0,2600**
rus	**-0,0669**	**0,4480**	**0,0449**	**-0,5709**	**-0,0831**	**-0,3385**
fins	**-0,0514**	**-0,2051**	**-0,0850**	**0,2733**	0,0117	**0,5113**
swe	-0,0335	**-0,1387**	-0,0017	**0,1018**	**0,1108**	**0,1055**

Tabel L1.3 Spearmani korrelatsioonikordajad MixFit algoritmi ja peakomponentanalüüs+diskriminantanalüüsi tulemuste vahel, 10 peakomponenti

	EST	LAT	RUS	FIN.S	SWE	FIN.N
est	**0,0726**	**-0,3211**	**-0,0677**	**0,4842**	**0,0591**	**0,2306**
lat	**-0,1096**	**0,4754**	-0,0187	**-0,4867**	**-0,0891**	**-0,2100**
rus	**-0,0583**	**0,3453**	**0,0826**	**-0,5149**	**-0,0592**	**-0,3242**
fins	**-0,0529**	**-0,1906**	**-0,0825**	**0,2553**	0,0095	**0,4887**
swe	-0,0340	**-0,1389**	-0,0041	**0,1025**	**0,0942**	**0,1217**

Tabel L1.4 Spearmani korrelatsioonikordajad MixFit algoritmi ja peakomponentanalüüs+diskriminantanalüüsi tulemuste vahel, 50 peakomponenti

	EST	LAT	RUS	FIN.S	SWE	FIN.N
est	**0,1348**	**-0,1477**	**-0,0426**	**0,2432**	0,0302	-0,0341
lat	**-0,1141**	**0,3529**	-0,0127	**-0,3539**	**-0,0630**	**-0,1320**
rus	**-0,0675**	**0,0602**	**0,1110**	**-0,2146**	-0,0136	**-0,1300**
fins	**-0,0577**	**-0,1940**	**-0,0809**	**0,2689**	0,0062	**0,4867**
swe	**-0,0465**	**-0,0691**	-0,0127	0,0288	**0,1005**	0,0191

Tabel L1.5 Spearmani korrelatsioonikordajad Toomas Halleri algoritmi ja peakomponentanalüüs+diskriminantanalüüsi tulemuste vahel, 100 peakomponenti

	EST	LAT	RUS	FIN.S	SWE	FIN.N
est	**0,1351**	**-0,1230**	**-0,0374**	**0,2040**	0,0324	**-0,0712**
lat	**-0,0987**	**0,3100**	-0,0014	**-0,3222**	**-0,0558**	**-0,1214**
rus	**-0,0664**	0,0322	**0,1124**	**-0,1938**	-0,0172	**-0,1042**
fins	**-0,0530**	**-0,1951**	**-0,0802**	**0,2669**	0,0108	**0,4769**
swe	-0,0275	**-0,0598**	-0,0208	0,0215	**0,0677**	0,0119

Lisa 2. Seosed fenotüüpide ja päritolukomponentide vahel

Järgnevates tabelites on välja toodud pikkusele ja igale päritolukomponendile vastavate korrelatsioonikordajate p-väärtused. Silmade- või juuksevärvi jaoks on p-väärtus vastava päritolukomponendi dispersioonanalüüsi mudeli olulisuse tõenäosus ning hinnang ühe faktortaseme keskmine päritolukomponendi kohta.

Tabel L2.1 Päritolukomponentide ja fenotüüpide seosed, 2 peakomponenti diskriminantanalüüsis

LDA2		est		lat		rus		fins	
		Hinnang	p-väärtus	Hinnang	p-väärtus	Hinnang	p-väärtus	Hinnang	p-väärtus
	Pikkus (korre-latsioon)	0,0584	<.0001	-0,0426	0,0002	-0,0428	0,0002	-0,0322	0,0051
Silma-värv (kesk.)	Hall (N=3303)	0,7962	0,0003	0,0050	0,0156	0,1842	0,0005	0,0145	0,1902
	Sinine (N=2224)	0,8145		0,0039		0,1667		0,1486	
	Pruun (N=835)	0,8201		0,0027		0,1709		0,0063	
	Roheline (N=1048)	0,8184		0,0038		0,1638		0,0140	
Juukse-värv (kesk.)	Blond (N=1701)	0,8183	0,0478	0,0039	0,5949	0,1615	0,0016	0,0163	0,4721
	Pruun (N=5085)	0,8046		0,0045		0,1786		0,0123	
	Must (N=549)	0,7963		0,0041		0,1872		0,1232	
	Punane (N=66)	0,8240		0,0021		0,1522		0,2162	

Tabel L2.2 Päritolukomponentide ja fenotüüpide seosed, 5 peakomponenti diskriminantanalüüsis

LDA5		est		lat		rus		fins	
		Hinnang	p-väärtus	Hinnang	p-väärtus	Hinnang	p-väärtus	Hinnang	p-väärtus
	Pikkus (korre-latsioon)	0,0432	0,0002	-0,0383	0,0009	-0,0359	0,0018	-0,0213	0,0635
Silma-värv (kesk.)	Hall (N=3303)	0,8312	0,0223	0,0034	0,0305	0,1602	0,0367	0,0052	0,4980
	Sinine (N=2224)	0,8448		0,0027		0,1477		0,0047	
	Pruun (N=835)	0,8466		0,0018		0,1501		0,0015	
	Roheline (N=1048)	0,8475		0,0025		0,1449		0,0051	
Juukse-värv (kesk.)	Blond (N=1701)	0,8554	0,0003	0,0025	0,3611	0,1350	<,0001	0,0071	0,0690
	Pruun (N=5085)	0,8341		0,0031		0,1587		0,0041	
	Must (N=549)	0,8244		0,0029		0,1722		0,0005	
	Punane (N=66)	0,8760		0,0009		0,1078		0,0154	

Tabel L2.3 Päritolukomponentide ja fenotüüpide seosed, 50 peakomponenti diskriminantanalüüsis

LDA50		est		lat		rus		fins	
		Hinnang	p-väärtus	Hinnang	p-väärtus	Hinnang	p-väärtus	Hinnang	p-väärtus
	Pikkus (korre-latsioon)	0,0253	0,0279	-0,0316	0,0060	0,0108	0,3475	-0,0227	0,0477
Silma-värv (kesk.)	Hall (N=3303)	0,9876	0,3754	0,0029	0,4084	0,0038	0,8692	0,0057	0,5822
	Sinine (N=2224)	0,98872		0,0022		0,0033		0,0058	
	Pruun (N=835)	0,9932		0,0008		0,0036		0,0022	
	Roheline (N=1048)	0,9869		0,0032		0,0046		0,0053	
Juukse-värv (kesk.)	Blond (N=1701)	0,9859	0,5794	0,0029	0,9243	0,0028	0,3213	0,0084	0,0543
	Pruun (N=5085)	0,9891		0,0025		0,0038		0,0045	
	Must (N=549)	0,9899		0,0025		0,0065		0,0012	
	Punane (N=66)	0,9846		0,0000		0,0000		0,0153	

Tabel L2.4 Päritolukomponentide ja fenotüüpide seosed, MixFit algoritmi tulemused

MixFit		EST		LAT		RUS		FIN.S	
		Hinnang	p-väärtus	Hinnang	p-väärtus	Hinnang	p-väärtus	Hinnang	p-väärtus
	Pikkus (korre-latsioon)	0,0298	0,0094	-0,0271	0,0185	-0,0038	0,7342	-0,0174	0,1297
Silma-värv (kesk.)	Hall (N=3303)	0,4887	0,1152	0,2309	0,0434	0,0719	0,0670	0,1274	0,0956
	Sinine (N=2224)	0,4925		0,2175		0,0747		0,1373	
	Pruun (N=835)	0,4748		0,2148		0,0890		0,1232	
	Roheline (N=1048)	0,5042		0,2093		0,0772		0,1327	
Juukse-värv (kesk.)	Blond (N=1701)	0,4918	0,6401	0,2121	0,2801	0,0729	0,5644	0,1455	0,0003
	Pruun (N=5085)	0,4914		0,2257		0,0748		0,1249	
	Must (N=549)	0,4796		0,2255		0,0844		0,1317	
	Punane (N=66)	0,5178		0,2240		0,0723		0,1221	

Lisa 3. Programmikoodid

Järgnevalt on välja toodud kasutatud R-i koodid.

20 000 suuremat varieeruvust andva SNP markeri leidmine

```
est<-t(est)
lat<-t(lat)
swe<-t(swe)
rus<-t(rus)
finn<-t(finn)
fins<-t(fins)
koik<-rbind(est,lat,rus,fins,finn,swe)
## rahvuse tunnus:
rahv<-rep(1:6,c(100,88,96,100,84,100))
## teen funktsiooni, mis arvutab skaleeritud tunnustele
## leitud grupikeskmiste standardh?lbe:
grsd<-function(x) sd(tapply(scale(x),rahv,mean))
s1<-apply(koik[,1:50000],2,grsd)
s2<-apply(koik[,50001:100000],2,grsd)
s3<-apply(koik[,100001:150000],2,grsd)
s4<-apply(koik[,150001:200000],2,grsd)
s5<-apply(koik[,200001:273454],2,grsd)
s<-c(s1,s2,s3,s4,s5)
s<-rev(sort(s))  # sorteeritakse suuremast väiksemani
plot(s)     #joonis 2.2.1
koik1<-koik[,names(s[1:20000])]  ## väiksem andmestik

#Nopime välja need SNP-d, mis on olemas ka andmestiku B inimestel
koik1=koik1[,c(colnames(genot))]
```

Peakomponentanalüüsi tegemine ning selle alusel prognoosimine

```
#Teostame peakomponentanalüüsi saadud skaleeritud andmetele
pc1<-prcomp(scale(koik1))

#Et arvutused ei läheks liiga mahukaks, jagame andmestiku, millele tahame
#saada prognoose, juppideks
genot1=genot[1:1000,]
genot2=genot[1001:2000,]
genot3=genot[2001:3000,]
genot4=genot[3001:4000,]
genot5=genot[4001:5000,]
genot6=genot[5001:6000,]
genot7=genot[6001:7000,]
genot8=genot[7001:7606,]

#Iga alamandmestiu puhul teostame sama skaleerimise nagu tehti andmestikule
#koik1:
kv=apply(koik1,2,mean)
sd=apply(koik1,2,sd)
```

```
genot1 <- t(t(genot1)-kv)
genot1<- t(t(genot1)/sd)
genot2 <- t(t(genot2)-kv)
genot2<- t(t(genot2)/sd)
genot3 <- t(t(genot3)-kv)
genot3<- t(t(genot3)/sd)
genot4 <- t(t(genot4)-kv)
genot4<- t(t(genot4)/sd)
genot5 <- t(t(genot5)-kv)
genot5<- t(t(genot5)/sd)
genot6 <- t(t(genot6)-kv)
genot6<- t(t(genot6)/sd)
genot7 <- t(t(genot7)-kv)
genot7<- t(t(genot7)/sd)
genot8 <- t(t(genot8)-kv)
genot8<- t(t(genot8)/sd)

#Liidame andmestikud jälle kokku ja teisendame maatriksiks:
genot=rbind(genot1,genot2,genot3,genot4,genot5,genot6,genot7,genot8)
genot=as.matrix(genot)

#Teeme saadud skaleeritud genotüübiväärtustele lineaarteisenduse eelnevalt
#arvutatud peakomponentanalüüsi alusel
tulemus=data.frame(genot%*%pc1$rotation[,1:568])

#Peakomponentide varieeruvuse kirjeldamine
summary(pc1)

#Joonis andmestiku A vaatlustest, kasutades esimest ja teist peakomponenti
plot(pc1$x[,2],pc1$x[,1],col=rahv)
legend("topleft",c("est","lat","rus","fins","finn","swe"),pch=1,col=1:6)
#Lisame andmestiku A vaatlustele prognoositud väärtused
points(tulemus[,2],tulemus[,1],col='7')

#Joonis andmestiku B vaatlused maakondade kaupa esimese kahe peakomponendi
#teljestikus.
library(ggplot2)
library(dplyr)
kesk_mk=tulemus%>%
  group_by(synnimk)%>%
  summarise(keskminePC1=mean(PC1),keskminePC2=mean(PC2))
kesk_mk$mk=c("","Hiiu","Harju","Ida-Viru","Jõgeva","Järva","Lääne","Lääne-
Viru","Põlva","Pärnu","Rapla","Saare","Tartu","Valga","Viljandi","Võru","XX
51","XX59","XX70","XX84")
ggplot(data=kesk_mk[kesk_mk$synnimk!=""&kesk_mk$synnimk!="XX51"&kesk_mk$syn
nimk!="XX59"&kesk_mk$synnimk!="XX70"&
kesk_mk$synnimk!="XX84",],aes(x=keskminePC2,y=keskminePC1))+geom_point(size
=2)+geom_text(aes(label=mk),hjust=0, vjust=0,size=5)
```

MixFit tulemuste andmestiku ning andmestiku C kombineerimine

```
f3=f3[f3$VCODE%in%row.names(f1),]
f4=f1[row.names(f1)%in%f3$VCODE,]
f4$VCODE=row.names(f4)
library(dplyr)
f3=arrange(f3,VCODE)
f4=arrange(f4,VCODE)
kokku=cbind(f3,f4[,1:18])
```

Teostame lineaarsed diskriminantanalüüsid, prognoosime ning vaatleme, kui hästi prognoosid korreleeruvad.

```
#LDA 2 peakomponenti
require(MASS)
comp2<-as.data.frame(pc1$x[,1:2])
ld <- lda(rahv~.,data=comp2)
#aposterioorsed tõenäosused
pr<- predict(ld, tulemus[,1:2])
lda2=data.frame(round(pr$posterior,5))

lda2=lda2[row.names(lda2)%in%kokku$VCODE,]
lda2$VCODE=row.names(lda2)
lda2=merge(lda2,kokku,by="VCODE")
cor(lda2[,c(2:5,7)],lda2[,c(8,11,12,10,13,9)],method="spearman")
#Korrelatsioonikordajate olulisuse tõenäosused:
library(Hmisc)
rcorr(as.matrix(lda2[,c(2:5,7,8,11,12,10,13,9)]),type="spearman")$P

#LDA 5 peakomponenti
comp2<-as.data.frame(pc1$x[,1:5])  #Siit valida mitu peakomponenti
ld <- lda(rahv~.,data=comp2)
#aposterioorsed tõenäosused
pr<- predict(ld, tulemus[,1:5])    #Siit valida mitu peakomponenti
lda5=data.frame(round(pr$posterior,5))

lda5=lda5[row.names(lda5)%in%kokku$VCODE,]
lda5$VCODE=row.names(lda5)
lda5=merge(lda5,kokku,by="VCODE")
#Spearmani korrelatsioonikordajad
cor(lda5[,c(2:5,7)],lda5[,c(8,11,12,10,13,9)],method="spearman")
#Korrelatsioonikordajate olulisuse tõenäosused:
library(Hmisc)
rcorr(as.matrix(lda5[,c(2:5,7,8,11,12,10,13,9)]),type="spearman")$P

#Analoogiliselt saab ka ülejäänud variandid kätte, võttes vastavalt rohkem
#peakomponente
```

```
#Karpdiagrammide joonistamine
pr<- predict(ld, tulemus[,1:2])
library(reshape2)

ggplot(melt(data.frame(round(pr$posterior,5))
),aes(x=variable,y=value))+geom_boxplot()+ylab("Tõenäosus")+xlab(" ")
ggplot(melt(f3[,c(4,7,10,6,5,11)]),aes(x=variable,y=value))+geom_boxplot()+
ylab("Tõenäosus")+xlab(" ")
```

Kasutatud SASi koodid

Toimingud MixFit algoritmi tulemustega.
Leiame korrelatsioonid pikkuste vahel.

```
proc corr data=kokku;
var pikkus EST LAT RUS FIN_S ;
run;
```

Eemaldame edasise vaatluse alt albiinod ja muu silmavärviga inimesed.

```
data kokku2;
set kokku;
run;

data kokku2;
set kokku2;
if silmvarv=5 then delete;
if silmvarv=6 then delete;
run;
```

Teeme uuritavatele rahvusgruppidele dispersioonanalüüsi mudelid silmavärvide kaupa.

```
proc glm data=kokku2;
class silmvarv;
model EST=silmvarv;
means silmvarv;
run;
proc glm data=kokku2;
class silmvarv;
model LAT=silmvarv;
means silmvarv;
run;

proc glm data=kokku2;
class silmvarv;
model RUS=silmvarv;
means silmvarv;
run;

proc glm data=kokku2;
class silmvarv;
model FIN_S=silmvarv;
means silmvarv;
run;
```

Teeme uuritavatele rahvusgruppidele dispersioonanalüüsi mudelid juuksevärvide kaupa.

```
data kokku3;
set kokku;
run;

data kokku3;
set kokku3;
if juuksevarv=5 then delete;
if juuksevarv=6 then delete;
run;


proc glm data=kokku3;
class juuksevarv;
model EST=juuksevarv;
means juuksevarv;
run;
proc glm data=kokku3;
class juuksevarv;
model LAT=juuksevarv;
means juuksevarv;
run;
proc glm data=kokku3;
class juuksevarv;
model RUS=juuksevarv;
means juuksevarv;
run;
proc glm data=kokku3;
class juuksevarv;
model FIN_S=juuksevarv;
means juuksevarv/tukey;
run;
```

Toimingud MixFit algoritmi tulemustega.
Leiame korrelatsioonid pikkuste vahel.

```
proc corr data=lda2;
var pikkus est lat rus fins ;
run;
```

Eemaldame edasise vaatluse alt albiinod ja muu silmavärviga inimesed.

```
data lda2_2;
set lda2;
run;

data lda2_2;
set lda2_2;
if silmvarv=5 then delete;
if silmvarv=6 then delete;
run;
```

Teeme uuritavatele rahvusgruppidele dispersioonanalüüsi mudelid silmavärvide kaupa.

```
proc glm data=lda2_2;
class silmvarv;
model EST=silmvarv;
means silmvarv;
means silmvarv/tukey;
run;
```

```
proc glm data=lda2_2;
class silmvarv;
model LAT=silmvarv;
means silmvarv;
means silmvarv/tukey;
run;

proc glm data=lda2_2;
class silmvarv;
model RUS=silmvarv;
means silmvarv;
means silmvarv/tukey;
run;

proc glm data=lda2_2;
class silmvarv;
model fins=silmvarv;
means silmvarv;
run;
```

Kõik ülejäänud mudelid viiele ja viiekümne peakomponendi korral leitakse analoogiliselt.

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Sven Erik Ojavee (sünnikuupäev 05.05.1993)

1. Annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose „Geneetiliste päritolukomponentide määramine mitmemõõtmelise statistika meetodite abil“, mille juhendajad on Krista Fischer ja Toomas Haller,
   1.1 reprodutseerimiseks säilitamise ja üldsusele kättesaadavaks tegemise eesmärgil, sealhulgas digitaalarhiivi DSpace-is lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;
   1.2 üldsusele kättesaadavaks tegemiseks Tartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.
2. olen teadlik, et punktis 1 nimetatud õigused jäävad alles ka autorile.
3. Kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest tulenevaid õigusi.

Tartus, 29.04.2015